DOCUMENT RESUME

ED 128 402 TM 005 585

TITLE Proceedings of the Annual Conference of the Military Testing Association (17th, Fort Benjamin Harrison, Indiana, September 15-19, 1975).

PUB DATE [Sep 75]

NOTE 799p.

EDRS PRICE MF-$1.50 HC-$42.19 Plus Postage.

DESCRIPTORS Aptitude Tests; Certification; Computer Assisted Instruction; Computer Oriented Programs; *Conference Reports; Course Evaluation; Educational Technology; Equivalency Tests; High School Graduates; Information Utilization; Job Analysis; Job Skills; Job Training; Measurement Techniques; *Military Personnel; Military Training; Performance Criteria; Performance Tests; *Personnel Evaluation; *Predictive Ability (Testing); Program Evaluation; Qualifications; Research Methodology; Screening Tests; Simulation; *Task Analysis; Technical Education; *Testing; Testing Problems; Test Validity; Vocational Aptitude; Vocational Counseling

IDENTIFIERS Applied Performance Testing; Armed Services Vocational Aptitude Battery; *Military Testing Association

ABSTRACT

The 62 papers presented at the 1975 conference of the Military Testing Association cover almost all areas of military and occupational assessment and evaluation, and are arranged in 19 "common subject-matter groupings": Symposium (on Aptitude Testing), Training Extension Courses, Computerized Testing, Task Validation and Qualification Standards, Computer Based Training, Testing/Assessment: Issues and Innovations, Screening and Counseling Enlistees, Occupational Analysis and Training Programs, Skill Qualification Under Enlisted Personnel Management System (EPMS), Research and Measurement Methodology, Demonstrating Occupational Competency, Uses of Occupational Analysis Data, Non-Cognitive Assessment, Task Analysis Training and Evaluation, Organizational Factors in Performance, Criterion Referenced Measurement, Assessing Job and Grade Requirements, and Measurement and Prediction. (BW)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED128402

PROCEEDINGS

17th ANNUAL CONFERENCE
of the
MILITARY TESTING ASSOCIATION

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

5-19 SEPTEMBER 1975

Coordinated by

US Army Enlisted Evaluation Center
Fort Benjamin Harrison, Indiana

TM005 585

PROCEEDINGS

17th Annual Conference
of the
Military Testing Association

Coordinated By:

U. S. Army Enlisted Evaluation Center
Fort Benjamin Harrison, Indiana 46249

Marriott Inn
Indianapolis, Indiana
15 - 19 September 1975

TABLE OF CONTENTS

PAGE

Officers and Committee Chairmen 1

Listing of Presentations by Common Subject-Matter Groupings (As they were presented) 2

Program Participants in Alphabetical Order 9

Papers in Order of the "Listing of Presentation by Common Subject-Matter Groupings" 15

Steering Committee Minutes 773

Minutes of Concluding General Session of MTA Conference 778

Military Testing Association By-Laws 780

List of Conferees 786

NOTE:

Due to space limitations, the editors have elected not to publish any of the appendixes which were included with many of the papers. Persons desiring these appendixes or any other information relevant to these papers should contact the author(s). The official addresses of all participants are listed at the back of this publication.

FOREWORD

The papers presented at the Seventeenth Annual Conference of the Military Testing Association reflect the opinions of their authors only and are not to be construed as the official policy of any institution, government, or branch of armed service.

OFFICERS AND COMMITTEE CHAIRMEN

17th Annual Conference of the MILITARY TESTING ASSOCIATION

Col George W. Orton
MTA President

Clay V. Brittain
Conference Chairman

William W. Wance
Program Chairman

John F. Finucane
Facilities Chairman

Deborah A. Felkins
Banquet Chairman

M. Anita Jones
MTA Secretary

COMMITTEE CHAIRMEN

LTC Peter G. Jones	Coordinating
Travis C. Glenn	Registration
Richard K. Edwards	Transportation
Fred B. Honn	Proceedings

LISTING OF PRESENTATIONS BY COMMON SUBJECT-MATTER GROUPINGS
(as they were presented)

Page

Keynote Address 15

SYMPOSIUM:

"Use of a Common Aptitude Test for Entry into all Military Services" . . . . Major Wayne S. Sellman 18

"Development of Armed Services Vocational Aptitude Battery, Forms 5, 6, and 7" . . . . Dr. Lonnie D. Valentine, Jr. 21

"Army Research in Support of the Goals of a Joint Services Vocational Aptitude Test Battery" . . . . Dr. M. A. Fischl 29
Mr. Robert M. Ross

"High School Use of ASVAB" . . . Mr. Harry D. Wilfong 37

TRAINING EXTENSION COURSES:

"TEC: A Manhattan Project in Educational Technology" . . . . CPT Wesley K. Roberts 43
Mr. Warren G. Lawson
CPT William D. Neal

"A Comparison of TEC and Conventional Training by Use of Performance Tests" . . . . . . Dr. John J. Kessler 58
Ms. Claramae S. Knerr
Mr. Ronald G. Downey

COMPUTERIZED TESTING:

"Discriminant Analysis Applied to Sequential Testing" . . . . . . Dr. David B. Thomas 70

Page

"Computerized Adaptive Ability Measurement". . . . . Mr. David J. Weiss 90

TASK VALIDATION & QUALIFICATION STANDARDS:

"Development of Qualification Standards and Evaluation Procedures-A Practical Approach". . . . . Mr. Alex L. Szaszy 128

"Delphi Applications to Job Task Analysis". . . . . Mr. John F. Kaczmarski 143

COMPUTER BASED TRAINING:

"Development and Validation of CII Performance Criterion Measures and Courseware" . . . . . Dr. W. G. Hoyt 153
Mr. A. K. Butler
Dr. D. J. Chesler

"CMI--Navy Style". . . . . Dr. Worth Scanland 183

TESTING/ASSESSMENT: ISSUES AND INNOVATIONS:

"Legal Policies Concerning Testing and Their Effect on Military Personnel". . . . . Ms. Darryl Lang 189

"One MOS Proficiency Test to Cover all Skill Levels". . . . .Mr. Herman A. Mahnen 198

"The New Army Enlisted Performance Rating Forms". . . . . Mr. Arthur G. Hermansen 207

SCREENING & COUNSELING ENLISTEES:

"Early Identification of Maladaptive Accessions". . . . . . Dr. Jeffrey E. Kantor
Ms. Nancy Guinn 221

"Comparative Performance of High School Graduates/Non-Graduates in Military Service". . Dr. Jeffrey E. Kantor 232
Ms. Nancy Guinn

"An Experimental Pre-Enlistment Screening Procedure". . . . . Mr. Ted Yellen 244

Page

"A Survey of the Developmental Research for Project CONTRACT (Computerized Navy Techniques for Recruit Assignment, Counseling, and Testing)". . . . . Mr. Bernard A. Rafacz 250

OCCUPATIONAL ANALYSIS AND TRAINING PROGRAMS:

"New CODAP Programs for Analyzing Task-Factor Information". . . . . Dr. Raymond E. Christal 266
Mr. Johnny J. Weissmuller

"A Universal Model for Evaluating Basic Electronic Courses in Terms of Field Utilization of Training". . . . . Dr. Walter E. Driskill 283
Major Thomas J. O'Connor
Mr. Hendrick W. Ruck
Mr. Karl A. Hickerson

"The Use of Occupational Analysis Outputs to Assist Training Design in the Royal Navy". . . . . CDR C. J. Nicholl 300

SKILL QUALIFICATION TESTING UNDER EPMS:

"The Army's Adoption of Performance Testing for MOS Proficiency". . . . . CPT Harry W. Porthouse 310

"Current and Projected Research in Skill Qualification Testing". . . . . . Dr. Douglas L. Young 326

"Issues in Skill Qualification Testing". . . . . . .Dr. Stephen F. Hirshfeld 339

RESEARCH & MEASUREMENT METHODOLOGY:

"PRIME: Personnel Research Instrumentalities for Measurement and Evaluation". . . . . . . Dr. William J. Moonan 344

"VARSEL: Variable Selection for Multiple-Purpose Prediction Systems Without Using External Criteria". . . . . Mr. R. Bruce Gould 363
Dr. Raymond E. Christal

Page

"Capturing Performance Rating Policies from Multitrait-Multirater Data". . . . . Major Jimmy L. Mitchell 377

DEMONSTRATING OCCUPATIONAL COMPETENCY:

"A Survey of Current Practices In Occupational Certification and Licensing Testing". . . . . .Mr. Michael D. Hiscox 388
Dr. Dean H. Nafziger

"Program Level Credit Awards for Occupational Competencies: Military-Civilian Linkages". . . . Dr. Duane O. Geiken 401

USES OF OCCUPATIONAL ANALYSIS DATA:

"Occupational Analysis - A Brief Introduction". . . . . . . Mr. B. Michael Berger 408

"The Use of CODAP to Define an Emergency Rating". . . . . . . .LT(JG) Allen R. Thuring 410

"Development of Procedures for Evaluating the U. S. Coast Guard's Post-Graduate and Post-Commission Training Program". . . . . .Mr. R. S. Lanterman 419
Mr. J. W. Cunningham

"Criterion-Referenced Personnel Assessment: An Overview". . . . . Mr. Lawrence S. Buck 423

"Evaluation of Manpower Policies Using Collected Occupational-Job Analysis Information" Dr. Raymond O. Waldkoetter 437
Mr. Peter F. Newton

"From Job Task Analysis to Behavioral Objectives By The Numbers". . . . . LCDR Raymond D. Bland 445

NON-COGNITIVE ASSESSMENT:

"Quality of Life in the U. S. Air Force". . . . . LTC T. Roger Manley 453
Major Charles W. McNichols
CPT Robert A. Gregory

Page

"Assignment Location Preferences of Airmen". . . . . Dr. Joe T. Hazel 468
Mr. William J. Stacy
Mr. Lloyd D. Burtch

"The Affective Domain and the Military Short Course". . . . .Major Ronald G. Swanson 485

SIMULATION AND APPLIED PERFORMANCE TESTING:

"The State of the Art in Applied Performance Testing". . . . . . . Mr. Thomas P. Sachse 492
Dr. James R. Sanders

"Sampling and Simulation in Tests of Job Performance". . . . . Ms. Elaine N. Taylor 508
Mr. Robert Vineberg
Mr. William C. Osborn

"A Criterion-Referenced Pilot Performance Measurement System". . . . . CPT John H. Fuller 512
Mr. Wayne L. Waag
Mr. Edward E. Eddowes
Mr. Robert R. Fuller

"The Decision Measurement System - Current Status". . . . . Mr. F. B. Braun 524
Dr. M. A. Robinson
Mr. J. E. Tindall, Jr.

TASK ANALYSIS TRAINING & EVALUATION:

"Task Analysis: The Basis for Criterion Test and Curriculum Design". . . . . .Mr. Robert N. Johnson 538
1LT James N. Richman

"Determining Training Priorities for Job Tasks". . . . . .LTC Donald F. Mead 551

"Task Factor Benchmark Scales for Use in Determining Training Priorities". . . . . Squadron Leader K. Goody 559
Squadron Leader W. J. Watson

"The Ultimate Catastrophe". . . . .LCDR Clarence L. Walker 569

Page

ORGANIZATIONAL FACTORS IN PERFORMANCE:

"Measurement of Unit Effectiveness: A Criterion Development Problem". . . . . Mr. Samuel Bowser 571

"Effect of Organizational Structure on Electronics Maintenance aboard Navy Ships". . . Mr. Herman L. Williams
Mr. Lloyd S. Standlee 587

"Human Performance in Total Ship Planning and Acquisition". . . Dr. James McGuinness
Mr. James P. Jenkins 591

CRITERION REFERENCED MEASUREMENT:

"A Bayesian Method for Maximizing Correct Mastery Classifications". . . . . Dr. Frederick Steinheiser, Jr. 604

"An Empirical Investigation of Four Criterion-Referenced Testing Models". . . . . Mr. Kenneth I. Epstein 627

ASSESSING JOB & GRADE REQUIREMENTS:

"Matching Job Education Requirements with Candidates' Educational Attainments - A Pilot Methodological Study". . . . . Squadron Leader William J. Watson
Squadron Leader Kenneth Goody 640

"Application of a Method for Determining Officer Grade Requirements". . . . . Dr. Gary N. Matthews
Mr. William J. Stacy 659

"Physical Demand Profiles for Four Airman Career Ladders". . . . Mr. Kenneth G. Koym 667

"A Method of Determining Desirable Task Experiences for First-Line Supervisors". . . . Mr. William J. Stacy
Dr. Joe T. Hazel 681

Page

"Interservice Procedures for Instructional Systems Development: Executive Overview". . . . . .Dr. Robert K. Branson 691
Dr. Gail T. Rayner
Mr. J. Lamarr Cox
Mr. John P. Furman

"Interservice Procedures for Instructional Systems Development: Implications for Army Training". . . . . .CPT Wesley K. Roberts 705

MEASUREMENT AND PREDICTION:

"An Investigation of the Air Force Officer Qualifying Test". . .Major Warren B. Shull 716

"Alternative Item Response Weighting Procedures: Development and Evaluation". . . . Mr. William A. Sands 732

"Prediction of Army Officer Performance". . . . . Dr. Arthur C. F. Gilbert 747

"The Current Predictive Validity of the Flight Aptitude Selection Tests". . . . . Mr. R. L. McMullen 758
Mr. R. F. Eastman

PROGRAM PARTICIPANTS IN ALPHABETICAL ORDER

	PAGE
BERGER, Mr. B. Michael Occupational Analysis - A Brief Introduction	408
BLAND, LTC Raymond D. From Job Task Analysis to Behavioral Objectives "By The Numbers"	
BOWSER, Mr. Samuel Measurement of Unit Effectiveness: A Criterion Development Problem	571
BRANSON, Mr. Robert K.(et al) Interservice Procedures to Instructional Systems Development: Executive Overview	691
BRAUN, Mr. F. B. (et al) The Decision Measurement System - Current Status	524
BUCK, Mr. Lawrence S. Criterion-Referenced Personnel Assessment: An Overview	423
CHRISTAL, Dr. Raymond E. (et al) New CODAP Programs for Analyzing Task-Factor Information	266
DRISKILL, Mr. Walter E. (et al) A Universal Model for Evaluating Basic Electronic Courses in Terms of Field Utilization of Training	283
EPSTEIN, Mr. Kenneth I. An Empirical Investigation of Four Criterion-Referenced Testing Models	627
FISCHL, Mr. M. A. Army Research in Support of the Goals of a Joint Services Vocational Aptitude Test Battery	29
FULLER, Mr. John H. (et al) A Criterion-Referenced Pilot Performance Measurement System	512

PAGE

GEIKEN, Dr. Duane O. 401
Program Level Credit Awards for Occupational Competencies: Military-Civilian Linkages

GILBERT, Dr. Arthur C. F. 747
Prediction of Army Officer Performance

GOODY, Squadron Leader Kenneth (et al) 559
Task Factor Benchmark Scales for Use in Determining Training Priorities

GOULD, Mr. R. Bruce (et al) 363
VARSEL: Variable Selection for Multiple-Purpose Prediction Systems Without Using External Criteria

HAZEL, Dr. Joe T. (et al) 468
Assignment Location Preferences of Airmen

HERMANSEN, Mr. Arthur G. 207
The New Army Enlisted Performance Rating Forms

HIRSHFELD, Dr. Stephen F. 339
Issues in Skill Qualification Testing

HISCOX, Mr. Michael D. (et al) 388
A Survey of Current Practices in Occupational Certification and Licensing Testing

HOYT, Mr. W. G. (et al) 153
Development and Validation of CII Performance Criterion Measures and Courseware

JOHNSON, Mr. Robert N. (et al) 538
Task Analysis: The Basis for Criterion Test and Curriculum Design

KACZMARSKI, Mr. John F. 143
DELPHI Application to Job Task Analysis

	PAGE
KANTOR, Mr. Jeffrey E. (et al) Comparative Performance of High School Graduates/Non-Graduates in Military Service	232
KANTOR, Mr. Jeffrey E. (et al) Early Identification of Maladaptive Accessions	221
KESSLER, Mr. John J. (et al) A Comparison of TEC and Conventional Training by Use of Performance Tests	58
KOYM, Mr. Kenneth G. Physical Demand Profiles for Four Airman Career Ladders	667
LANG, Ms. Darryl Legal Policies Concerning Testing and Their Effect on Military Personnel	189
LANTERMAN, Mr. Richard S. (et al) Development of Procedures for Evaluating the U. S. Coast Guard's Post-Graduate and Post-Commission Training Program	419
McGUINNESS, Dr. James (et al) Human Performance in Total Ship Planning and Acquisition	591
McMULLEN, Mr. R. L. (et al) The Current Predictive Validity of the Flight Aptitude Selection Tests	758
MAHNEN, Mr. Herman A. One MOS Proficiency Test to Cover All Skill Levels	198
MANLEY, Mr. T. Roger (et al) Quality of Life in the U. S. Air Force	453
MATTHEWS, Mr. Gary N. (et al) Application of a Method for Determining Officer Grade Requirements	659
MEAD, LTC Donald F. Determining Training Priorities for Job Tasks	551

PAGE

MITCHELL, Major Jimmy L. 377
Capturing Performance Rating Policies from Multitrait-Multirater Data

MOONAN, Dr. William J. 344
PRIME: Personnel Research Instrumentalities for Measurement and Evaluation

NICHOLL, CDR C. J. 300
The Use of Occupational Analysis Outputs to Assist Training Design in the Royal Navy

PORTHOUSE, CPT Harry W. 310
The Army's Adoption of Performance Testing for MOS Proficiency

RAFACZ, Mr.Bernard A. 250
A Survey of the Developmental Research for Project Contract (Computerized Navy Techniques for Recruit Assignment, Counseling, and Testing)

ROBERTS, CPT Wesley K. (et al) 43
TEC: A Manhattan Project in Educational Technology

ROBERTS, CPT Wesley K. 705
Interservice Procedures for Instructional Systems Development

SACHSE, Mr. Thomas P. (et al) 492
The State-of-the Art in Applied Performance Testing

SANDS, Mr. William A. 732
Alternative Item Response Weighting Procedures: Development and Evaluation

SCANLAND, Dr. Worth 183
CMI-Navy Style

SELLMAN, Major Wayne S. 18
Use of a Common Aptitude Test for Entry into All Military Services

PAGE

SHULL, Major Warren B. 716
An Investigation of the Air Force Officer Qualifying Test

STACY, Mr. William J. (et al) 681
A Method of Determining Desirable Task Experiences for First-Line Supervisors

STEINHEISER, Dr. Frederick, Jr. 604
A Bayesian Method for Maximizing Correct Mastery Classifications

SWANSON, Major Ronald G. 485
The Affective Domain and the Military Short Course

SZASZY, Mr. Alex L. 128
Development of Qualification Standards and Evaluation Procedures - A Practical Approach

TAYLOR, Ms. Elaine N. (et al) 508
Sampling and Simulation in Tests of Job Performance

THOMAS, Dr. David B. 70
Discriminant Analysis Applied to Sequential Testing

THURING, LT(JG) Allen R. 410
The Use of CODAP to Define an Emergency Rating

VALENTINE, Mr. Lonnie D., Jr. 21
Development of Armed Services Vocational Aptitude Battery, Forms 5, 6, and 7

WALKER, LCDR Clarence L. 569
The Ultimate Catastrophe

WALDKOETTER, Dr. Raymond O. (et al) 437
Evaluation of Manpower Policies Using Collected Occupational-Job Analysis Information

PAGE

WATSON, Squadron Leader William J. (et al)
Matching Job Education Requirements with Candidates' Educational Attainments - A Pilot Methodological Study 640

WEISS, Mr. David J.
Computerized Adaptive Ability Measurement 90

WILFONG, Mr. Harry D.
High School Use of ASVAB 37

WILLIAMS, Mr. Herman L. (et al)
Effect of Organizational Structure on Electronics Maintenance Aboard Navy Ships 587

YELLEN, Mr. Ted 244
An Experimental Pre-Enlistment Screening Procedure

YOUNG, Dr. Douglas L. 326
Current and Projected Research in Skill Qualification Testing

KEYNOTE ADDRESS

Summary of General Gard's Address to MTA 1975

The Armed Forces today are faced with the unprecedented challenge of maintaining a sizeable Active and Reserve Component force without the benefit of selective service. This challenge comes during a period of social change so rapid and pronounced that its composite effect often is called a social revolution. One result of our affluent society has been degradation of the work ethic which in the past formed so significant a part of our social fabric. Today people are not willing to be thought of simply as factors of production. They insist upon opportunity for personal and professional growth. In addition to the desire to be treated well, people want their talents to be employed effectively. For some 30 years, under the selective service system, we were able to emphasize even low priority organizational objectives with little consideration of our personnel and their development. We must accommodate a reasonable response to the legitimate expectations of our soldiers.

In the Army, we have embarked on what we call EPMS, the Enlisted Personnel Management System. We are redesigning our Career Management Fields to provide advancement in each from the grade of E-1 to E-9. We are eliminating and combining some MOS at the lower grades and merging others at the higher grades. The purpose is to provide logical paths of career development for our soldiers from entrance into the Army to retirement.

Even in a buyer's market, we are having difficulty with criteria for admission into the Army. One problem concerns soldiers who cannot read but nevertheless are able to pass the required tests and get into the Army. Perhaps we have designed the tests so poorly that it requires little more than random chance to qualify.

For basic training, we have a "SMART Book," which some trainees can't read, outlining all the things the soldier has to be able to do in order to graduate. We became preoccupied with precise procedural criteria for end-of-course testing coincident with a reduction in the period allocated for basic training. The result of too much emphasis on what could be tested easily was that we were turning out civilians in uniform who were technically competent in a limited number of tasks but who were not socialized into the Army.

There are difficulties also with Advanced Individual Training. In a course for supply specialists, for example, 48 percent of the students failed. Investigation disclosed that the reading level required to understand the materials was nearly the equivalent expected of a college graduate.

Failure rates in other courses have run as high as 60 percent--and this in spite of the fact that we are receiving youngsters whose mental categories are relatively high. I suggest we have considerable work to do in our designs for evaluation as well as training.

The Army is now in the process of changing from a norm-referenced to a criterion-referenced approach to the evaluation of the occupational skills of its personnel. The purpose is to promote more realistic and effective training, but the new system also must continue to provide reliable and valid information regarding relative individual proficiencies for personnel management purposes. The Skill Qualification (SQT), to be based on essential tasks, will be the primary tool for acquiring this information. The Training and Doctrine Command (TRADOC) has established the Training Management Institute to bring together the talent necessary to design a methodology that will draw on such technology as CODAP, the Comprehensive Data Analysis Program, and the efforts of such agencies as the Army Research Institute for the Social and Behavioral Sciences. One significant area will involve how to determine what tasks are critical and how to test them. Another will be how to administer the SQT in different locations in a manner perceived by the soldier as equitable.

We find ourselves in the unaccustomed position of having to be selective in determining who can reenlist. This presents the problem of evaluating potential for success in positions of higher responsibility as well as skill qualification and past and current duty performance.

There often are inherent conflicts between the needs of program managers and of researchers who undertake to study or evaluate the programs. A research study on the Army's drug abuse program provides an example. About 3 years ago, we contracted for an expensive research project to assess the comparative effectiveness of alternative educational approaches to achieve certain objectives with different target groups. Only recently have the results of the study become available. The conclusion of the researchers was that drug education does not prevent drug abuse. We emphasized at the initial planning conference that we already knew this, and saw no use in rediscovering it. We had something entirely different in mind.

It seems to me that there are opportunities for significant improvement in our testing and evaluation. We have much to learn about the problems our soldiers experience in the Army; far too many are being released prior to completing their first terms of service. I have never seen a credible evaluation of any of our various human and social service programs. In the area of equal employment opportunity, we continue to be plagued by problems of cultural bias at work in our methods for selecting and evaluating the qualifications of our people. We have only begun to examine the issues involved in integrating women into the Army. We are starting to employ what are called organizational development techniques which offer high potential for the use of survey and feedback techniques along with other evaluation methodology to assist the commander. In addition, the Freedom of Information and Privacy Acts may complicate our problems further.

In looking at the program of this MTA Conference, I am tempted to paraphrase Dean Acheson and say that you may be trying to escape difficult problems by finding refuge in impossible ones. But I am pleased to see that you are addressing a number of the subject areas which are so important to the Armed Services today as we seek to maximize the employment of our human resources in achieving organizational effectiveness.

SYMPOSIUM

USE OF A COMMON APTITUDE TEST FOR ENTRY INTO ALL MILITARY SERVICES

Introductory Remarks

Major Wayne S. Sellman
Air Force Military Personnel Center

The Armed Services use a variety of tests for selecting and classifying their initial manpower accessions. Each Service test battery contains several sub-tests which are similar in content to tests in the other Service batteries, although the tests differ in format, length, difficulty pattern, or other characteristics.

In early 1966, the Assistant Secretary of Defense (Manpower and Reserve Affairs) directed the Services to explore the feasibility of a common aptitude test battery which could serve as an instrument for counseling high school students on vocational choices, could provide appropriate military service qualification data, and could be used in making job classification decisions about military enlistees. A working group, consisting of personnel test experts from all the Military Services, was formed to study the feasibility of such a test battery and to develop a prototype. The Armed Services Vocational Aptitude Battery (ASVAB) was developed from that effort.

ASVAB combined the best and most desired features of all the military classification tests and gave the DOD High School Testing Program a product usable by all the Services. The genesis of the High School Testing Program dates back to 1958, when some Air Force recruiters conceived the idea of administering a vocational aptitude test in the high schools, enabling them to earlier and more easily identify students with necessary aptitudes as potential recruits. With the success of the Air Force program the other Services initiated similar efforts. The resultant competition caused a certain degree of confusion and projected less than a professional image of the military. Thus, in 1966, the Department of Defense directed that the Services combine their high school testing activities and the DOD High School Testing Program was born.

The concept of a combined high school testing effort was good, but different parts of the action were assigned to the various Services. For example, the Army was tasked to develop the test while the Navy was requested to provide all counselor's materials. At the same time, the Air Force supervised the test scoring and return of the test results to the high schools. Recruiters went back into the schools, all using the same test, but still with their parochial interests at heart. Thus, ASVAB became an Army test, a Navy test, or an Air Force test depending upon the representative who got there first.

In 1973, the Armed Forces Vocational Testing Group (AFVTG) was formed in answer to this problem. Staffed by personnel from all the Services, the Group provides a single manager and interpreter of ASVAB for the DOD High School Testing Program. Dealing with the local recruiter through an interservice recruitment committee (organized by geographical location and representing each Service) the AFVTG assures equitable and professional administration of ASVAB.

The mission of the AFVTG is to place ASVAB in the schools, assure its currency and validity, and provide the student and counselor with assistance and information. The AFVTG does not recruit students. If the student is interested in the Military, the local recruiter provides him/her with information after the test has been administered. Improved recruiting, however, is a very important by-product of the program since the AFVTG does provide a list of student names and test scores to each of the Service recruiting commands. It should be noted that schools and students are aware of this recruiting advantage and of course that their participation in the program is strictly voluntary.

Concurrently with these efforts in the high school arena, the Office of the Assistant Secretary of Defense (Manpower and Reserve Affairs) was also considering the use of a single test across all Services for enlisted accessions. This single test notion was not a new one however. In 1951, the Armed Forces Qualification Test (AFQT) was introduced into the Services for just a role. For seven years, the AFQT served as the sole military enlistment test. However, in 1958 because of differences in enlistment/classification requirements, the Services convinced Defense that they should be authorized the use of supplementary aptitute tests along with the AFQT to refine the enlistment decision. This was followed in 1972 by still another Defense authorization - this time for Services to use selection tests of their own choice provided conversion tables to the AFQT were established.

Aware of and sensitive to this history of enlisted accessions testing, in April 1974, the Defense Manpower Policy Council, following the recommendation of the Assistant Secretary of Defense (Manpower and Reserve Affairs), decided to come full cycle and again use a single enlisted selection and classification battery for all Armed Forces. Since it was already in use within the DOD High School Testing Program, ASVAB was selected to be the single test with implementation scheduled for January 1976. The use of a common test was based on the notion that it would streamline the military procurement process and be advantageous to individuals, the Services, and to the Department of Defense. More specifically, a common enlistment test would insure that: (1) the high school testing program would be more useful to all Services, (2) applicants would not be subjected to multiple testing, (3) inter-service referrals would be facilitated, (4) the test development work of psychologists would be concentrated on a single test, and (5) it would form the basis for accurate cross-service comparisons which would facilitate Defense policy decisions and management. An ASVAB Steering Committee, with membership at the two-star level from Service personnel plans offices was chartered by Defense to oversee the completion and implementation of ASVAB as the single enlisted entrance test. The Air Force was designated as DOD Executive Agent.

In this symposium, we will discuss the development and implementation of ASVAB. Dr. Lonnie Valentine, Air Force Human Resources Laboratory, will describe the design of the test as well as share some of his experiences in building a test for Services with differing procurement policies, selection standards, and classification philosophies. He will be followed by Dr. Mike Fischl, Army Research Institute for the Behavioral and Social Sciences, who will present some data relative to the cross-service utility of ASVAB. Next, Dr. Harry Wilfong, Armed Forces Vocational Testing Group, will update information on the use of ASVAB in the high schools and make recommendations regarding how the high school and Service production testing programs can be made more compatible. Finally, Ms Jeannie Fites, Defense Manpower Research and Data Analysis Center, will comment on all papers with special emphasis placed on Department of Defense perspectives of problems encountered in managing and implementing such a joint Service project. Upon completion of this structured portion, we will be pleased to open up the symposium for comments or questions from the floor.

Development of Armed Services Vocational Aptitude Battery, Forms 5, 6, and 7

Lonnie D. Valentine, Jr.
Air Force Human Resources Laboratory

For a number of years, the Air Force used the Airman Qualifying Examination in various forms and revisions for classification and initial assignment of enlisted personnel. This examination was a short (about two hour) substitute for the longer Airman Classification Battery, and could be administered at field locations rather than at the basic training center. Its development and use was necessitated by the development of a recruiting plan under which recruits would be routinely targeted against assignment into one of four broad job clusters (Mechanical, Administrative, General, or Electronic), with final allocation into specific training or job slots within the cluster during basic training being predicted on consideration of level of aptitude in the cluster, individual desires, and available jobs.

In 1962 the Air Force Recruiting Service inaugurated a high school testing program using the AQE as the test instrument. Score information from the test was provided to school counselors for their use, and the program also served to identify to recruiters promising enlistment prospects from among graduating high school seniors.

After inauguration of the Air Force program, the other services also inaugurated high school testing programs. Partly as a function of high school complaints about competing programs, DoD directed that a joint service high school testing program be set-up, and in early 1966 a working group was established to develop the test instrument for the joint program. Development of the initial battery has been reported previously at MTA and is described in an ARI Technical Report; of interest here is the fact that Forms 1 through 4 of the battery consisted of nine scales which were common to the separate service classification batteries or were required for an AFQT score.

In early 1974 DoD directed that the services move expeditiously toward use of a common aptitude battery for enlistment production. OASD(MR&A) guidance was that ASVAB be redesigned to satisfy enlistment production requirements of all the services with high school usage being a secondary consideration. At the present time, operational implementation of the reconfigured battery is targeted for 1 Jan 76 within all services with testing centralized at the AFEES.

The remainder of this paper describes the battery as reconfigured for joint service enlistment production, provides rationale for the content, and provides some psychometric data on the battery. These data will not be as complete as had been originally planned because normative tryouts of the battery are still in progress and the analyses have not been completed.

Finally, I will discuss problems in battery development when so broad a range of goals must be satisfied and will briefly describe future plans.

Starting point for battery redesign was consideration of content of current service classification batteries and the uses being made of this content. In addition, near term plans of the services with respect to battery modifications were considered. From these reviews, a preliminary battery plan was developed at AFHRL for review at the other service labs.

In the course of a couple of subsequent joint service working group meetings, modifications to enhance the battery's acceptability to the services and to shorten testing time requirements were agreed upon.

The preliminary plan called for two perceptual tests, 12 cognitive power tests, and a rather lengthy Interest Inventory to include materials from the Army Classification Inventory, the Navy Vocational Interest Inventory, and the Air Force's Vocational Interest Choice Examination. It was estimated that the battery defined in the initial plan would require over four hours actual testing time. Table 1 summarizes content of the preliminary plan.

Table 1

Preliminary Plan for ASVAB*

Content Area		Number of Items
Attention to Detail		30
Numerical Operations		50
Word Knowledge		25
Arithmetic Reasoning		25
Space Perception		25
Mathematics Knowledge		25
Electronics Information		25
Radio Information		15
Mechanical Comprehension		25
Automotive Information		25
Shop Information		25
Biological Science		15
Physical Science		15
General Information		20
Interest Inventory		527**
Army Classification Inventory	(87)	
Navy Vocational Interest Inventory	(190)	
Vocational Interest Choice Exam	(250)	

---

* Estimated Testing Time = 4 hrs 6 min

** Total item pool before consolidation, where possible, of content from the three source inventories.

The clerical speed measure contained in earlier forms of the ASVAB was coding speed; this scale had been contained, but was later dropped from the ACB; it was quite different from clerical speed scales used in previous Air Force and Navy batteries. Army data suggested that Attention to Detail was preferable to coding speed for the clerical composite, and Air Force data showed lower validity for the ASVAB Administrative Composite than had been obtained from the AQE. (The difference lay in coding speed in ASVAB vs Numerical Operations in AQE).

Numerical Operations involves rapid addition, subtraction, multiplication, and division. It has demonstrated Air Force validity for the Administrative jobs; moreover, Navy had expressed interest in evaluating it for Navy schools.

Word Knowledge, Arithmetic Reasoning, and Space Perception have a long history as highly valid components of service selection and classification test batteries and, typically, constitute the service qualification composite, or AFQT.

Mathematics Knowledge was included because such a test is part of the present Army and Navy batteries; it is, essentially, an algebra and geometry test.

In the electronics area, two separate scales were proposed; an Electronics Information scale much like that in the present ASVAB and a Radio Information scale. Electronics Information has been included in service batteries over a number of years; in addition, Navy uses Radio Information items in their battery. While one would expect relatively high correlation between these scales, Navy gains some unique utility from their Radio Information items.

Mechanical Comprehension, Shop Information, and Automotive Information have typically formed the nucleus for Mechanical and Crafts Composites in the service batteries. Mechanical Comprehension deals with understanding of mechanical principles and movements while Automotive Information measures information obtainable through mechanical (especially automotive) experience; and Shop Information deals with repair procedures and tool usage for a variety of hands on activities.

Two separate science scales were proposed: Biological Science and Physical Science. The ACB contains a biological science scale which has demonstrated validity in the Army system, especially for medical specialties, and the Navy Battery contains a Physical Science scale which has demonstrated validity for some of their specialties. The initial proposal was to include these two areas, in equal numbers of items, in a General Science scale which could yield separate Biological and Physical Science scores.

Army's General Information scale is considered essential to their purposes, and its inclusion was also proposed; it is characterized by items dealing with sports, outdoor activities, weapons, and automobiles.

Remaining elements of the preliminary proposal were interest materials. Army uses an 87-item classification inventory, yielding four scores. These are used in formation of six Army composites. ACI is viewed by Army as essential since it provides differential variance to Army composites. Air Force and Navy labs have both also been engaged for some time in interest measurement research. Navy has developed the Navy Vocational Interest Inventory (NVII) for use in recruit counseling re job assignments, while Air Force has developed the Vocational Interest Choice Examination (VOICE) for similar purposes. These latter two instruments are quite different in format and measurement approach.

Preliminary proposal was to form a two-part interest inventory from these three instruments, thus providing a common interest data base for further analysis and refinement into a joint interest instrument.

Table 2

Final Plan for ASVAB

Content Area	Number of Items	Test Time (In minutes)
Attention to Detail	30	5
Numerical Operations	50	3
Word Knowledge	30	10
Arithmetic Reasoning	20	20
Space Perception	20	12
Mathematical Knowledge	20	20
Electronics Information*	30	15
Mechanical Comprehension	20	15
Automotive Information	20	10
Shop Information	20	8
General Science**	20	10
General Information	15	7
Classification Inventory***	87	20
	Total =	2 hrs 35 min

---

* Composed of 15 Electronics Information and 15 Radio Information items.

** Composed of 10 Biological and 10 Physical Science items.

*** Army items only.

Table 2 shows final content configuration for the battery. It was essential that the battery be shortened from its preliminary testing time estimates, especially for application in the joint services high school program; in addition, the various recruiting service commanders desired a test battery considerably shorter than the four hours required in the preliminary plan. Content shown in Table 2 was arrived at after a series of joint service committee deliberations, and was possible only because of various compromises from what would have been considered optimal. Note that, for the most part, scales in the final plan were shortened from the originally planned number of items, and one must expect some reduction in reliability as a result. Word Knowledge was the one exception to this general length reduction; this is because Coast Guard uses Word Knowledge to screen personnel for officer programs and believed that anything less than 30-items would be inadequate for that purpose.

The proposed Biological and Physical Science scales were combined into a single 20-item General Science scale with 10 items of each type. In like manner, the proposed Electronics Information and Radio Information scales were combined into one 30-item scale with 15 items of each type. All interest material, with the exception of the 87 items keyed for Army composites, was excluded; thus Navy and Air Force use of interest scales for counseling must be handled as separate, additional testing. The present 2 hour 35 minute time does not include test instructions and identification information entry time. When one allows a 25-30 minute estimate for these, testing exceeds the 2 hour 45 minute maximum limit often cited for most high schools.

A pool of 2700 test items (exclusive of Attention to Detail, Numerical Operations, and Interests) was assembled for experimental tryout and item analysis preparatory to final item selection. The excluded areas were handled as separate scale development efforts since one of them depended solely on use of empirically keyed interest items and the other two test ability at easy tasks under highly speeded conditions. The item pool was administered to samples of 500-600 cases at basic training centers and at AFEES; AFEES testing was necessary to obtain representation of cases rejected for enlistment. Item difficulty estimates were based on the full range mobilization population, and care was taken to include representation of women and ethnic minorities in the item analysis samples.

Table 3

Summary of ASVAB-5, 6, & 7 Item Difficulties*
(p values)

	Form 5		Form 6		Form 7	
Content Area	Aver. p	p range	Aver. p	p range	Aver. p	p range
General Information	.62	.39-.90	.61	.38-.90	.62	.44-.90
Word Knowledge	.61	.32-.86	.61	.32-.90	.61	.31-.90
Arithmetic Reasoning	.60	.27-.92	.61	.30-.94	.61	.30-.95
Space Perception	.60	.30-.90	.61	.29-.85	.61	.25-.83
Math Knowledge	.50	.33-.90	.50	.27-.90	.50	.31-.85
Electronics Information	.49	.27-.82	.50	.32-.85	.50	.28-.85
Mechanical Comprehension	.53	.30-.83	.53	.28-.87	.53	.33-.83
General Science	.56	.32-.85	.54	.32-.86	.55	.37-.90
Shop Information	.55	.30-.93	.57	.34-.90	.55	.34-.89
Automotive Information	.54	.34-.86	.53	.31-.81	.54	.33-.84

---

* Data on Attention to Detail and Numerical Operations are excluded since speed on the scale, rather than individual items, is critical. Classification Inventory is excluded since it is empirically keyed against Army Tech courses.

Table 4

Median ASVAB Item Internal Consistency Indexes*

Content Area	Form 5	Form 6	Form 7
General Information	.54	.48	.52
Word Knowledge	.42**	.42**	.40**
Arithmetic Reasoning	.40**	.39**	.41**
Space Perception	.42**	.41**	.42**
Math Knowledge	.57	.56	.54
Electronics Information	.48	.47	.46
Mechanical Comprehension	.54	.53	.53
General Science	.49	.53	.52
Shop Information	.56	.55	.56
Automotive Information	.61	.60	.62

---

* Data on Attention to Detail and Numerical Operations are excluded since speed on the scale, rather than individual items, is critical. Classification Inventory is excluded since it is empirically keyed against Army Tech courses.

** Phi coefficients, other indexes, or biserials.

Three forms of the battery (5, 6, and 7) were developed from these items. Table 3 summarizes item difficulty for these three forms, and Table 4 summarizes internal consistency data for them. The p values given in Table 3 are based on an AFQT stratified sample and are, thus, estimates for the mobilization population. All p values are uncorrected for chance success; it was observed that most subjects attempt all items. It might be noted that average p's for Word Knowledge, Arithmetic Reasoning, and Space Perception are about .61. These three scales yield ASVAB's AFQT composite which is used for selection; easy items are needed in these scales to allow for selection. Remaining scales are used in classification composites; an average p of .5 was sought for these. The one exception is General Information which is used only by Army and for which Army desired easy items.

With respect to internal consistency, all items selected for inclusion in the scales correlated significantly with total item type score and had misleads which correlated negatively with total item type score; most misleads demonstrated significant negative correlation.

At the present time, experimental administration of final versions of ASVAB 5, 6, and 7 for standardization and other psychometric determinations is in progress; these data will be reported soon in a Technical Report on battery development.

This round of joint service production battery development has identified a number of problems requiring special attention; some of these problems are generated by differences among service systems for selection and initial assignment of new personnel while others are attributable to situational factors. The most critical problems involve balance between desired battery length and differing requirements (both with respect to content and standards) of the several systems in which the battery must be used. While there is an appreciable "common" core among the various service and high school program content requirements (1) there are some unique service content requirements, (2) some of these unique requirements are not entirely appropriate for the high school program, and (3) their inclusion in the battery extends testing time beyond the limits that many schools will accommodate. Of equal criticality is the fact that service classification standards differ to some extent from each other; the impact of these standards differences is that longer, rather than shorter, paper-and-pencil scales are needed to provide sensitive cut points at a number of levels.

With centralization at AFEES of enlistment production testing and the emergence of adaptive or tailored testing as a viable alternative to traditional paper-and-pencil batteries, the services can circumvent these problems; however, adaptive procedures currently are not feasible for the high school program; moreover, a systematic development program is necessary before adaptive strategies could be implemented at AFEES. Meanwhile, time constraints will necessitate short ASVAB scales and will preclude implementation of Navy and Air Force interest inventories for pre-enlistment classification counseling.

A brief digression on this point is in order. The services have developed guarantee assignment recruiting strategies and are well along forward computer-based central assignment systems linked to the AFEES. A logical extension of this would provide for computer-linked tailored testing. This would, in turn, bypass the necessity for manual imput of test data into the assignment system.

Other problems in joint service production testing center around the need of those administering the program for multiple versions of the battery to guard against test compromise. This creates particular problems from a test development standpoint in that each version of the battery involving item changes requires restandardization; since deactivation of the cases available at AFEES for experimental testing are limited, their retention for an extra three or four hours of experimental testing creates logistics problems (e.g., subjects must be housed over night, test facilities are over worked, etc.). Development of three forms per year would place a near continuous load of this kind of AFEES.

A most viable solution to these test revision problems lies in the reintroduction of a battery in which each scale is developed and normed separately. This would greatly simplify AFEES normative support since a minimum of experimental time at AFEES for any one subject would be required. Usually, a maximum of 25 to 30 minutes experimental time would be required; if this were provided routinely, crash normative studies would be unnecessary.

Individual revised scales could be normed using their ASVAB predecessor as a reference. To assure adequacy of a revised scale, it would only be necessary to assure high correlation with the predecessor scale and comparable correlation with the other scales. From a generalized ASVAB intercorrelation matrix, standard composite conversions could be generated. These would be predicated on prior conversion of scales to a normal standardized metric. This approach would vastly multiply the potential number of battery forms. For example, three forms of each scale could yield over one and a half million forms of the battery (or combination of scales). If the battery were published as four booklets with three forms each, this would yield 81 forms of the battery achieved by different combinations of forms for the four booklets.

ARMY RESEARCH IN SUPPORT OF THE GOALS OF A JOINT SERVICES VOCATIONAL APTITUDE TEST BATTERY

M. A. FISCHL AND ROBERT M. ROSS
U. S. ARMY RESEARCH INSTITUTE FOR THE
BEHAVIORAL AND SOCIAL SCIENCES

Army policy supports Department of Defense instructions that the military services explore the feasibility of developing a single test battery to be used by all for initial qualification and vocational classification decisions. The test battery in use by the Army for these purposes at the time of the instruction was a comprehensive battery spanning general ability, specific aptitudes, and vocational interests. This battery enabled determination of qualification for the service, and differential classification of the entire enlisted intake into suitable training for in excess of 400 jobs, with most validity coefficients in the 0.6 to 0.7 range. Policy stipulated that the new battery should do no less for the Army.

As the specifications for the new battery were being developed, certain similarities and important differences were noted among the tests in use by the various services. To guide subsequent design thinking, the interest inventories of the Air Force and Navy were each administered to a sample of new Army enlistees to enable comparison of scores with Army interest test scores on these men and women. Similarly, the Navy's Electronics Technician Selection Test, which contains mathematics, physical science, and electrical information items, was administered to an Army sample to determine its comparability with analagous measures in the Army battery. Finally, a speed test of simple arithmetic was administered to an Army sample for purposes of evaluating its usefulness as a new clerical selector.

METHOD

The Navy Vocational Interest Inventory (NVII), the Air Force Vocational and Occupational Choice Examination (VOICE), the Navy Electronics Technician Selection Test (ETST), and an experimental Arithmetic Skill test were administered to a national sample of US Army enlisted men and women at three Reception Stations during September and October 1974. The stations were located in the Southeast, Midwest, and Far West. All examinees had previously been administered the Army Classification Battery (ACB) at examining stations, hence ACB scores were available. The two interest measures and the Arithmetic Skill test were administered each to 250 enlisted men and 75 enlisted women, the ETST to double those numbers.

The interest measures and ETST are well-known. The experimental Arithmetic Skill test requires description. This test taps attention to, and performance of, extremely simple arithmetic under a heavy speed loading. The items consist entirely of the addition or subtraction of two-digit numbers with no carrying. The top number is larger than the bottom so that there is no differential cue to addition or subtraction provided, and there is a random distribution of 48 of each type throughout the test. It is administered with a four-minute time limit, and the intention was to evaluate its potential as an indicator of clerical skill, not quantitative ability.

Two cautions should be noted concerning the sampling procedure employed. First, there were no cases from the lowest decile of the population on general ability, and there were fewer than half of the expected number of cases in the next lowest decile (percentiles 10-19). These restrictions were concomitants of the testing of enlistees,[1] instead of applicants, at a time when the minimum AFQT enlistment standard was at the 16th percentile. Second, in addition to general ability restriction, all examinees had previously qualified at the 30th percentile (Army Standard Score 90) in at least one aptitude area. The aptitude area composites contain general ability measures, measures of specific aptitudes, and measures of vocational interest. All correlation coefficients reported would be higher if the full range of general and specific abilities had been represented in the sample.

RESULTS

Table 1 presents the correlation of scales from the NVII and VOICE with their apparent counterparts, based on scale names, in the Army Classification Battery.

Table 2 presents intercorrelations among ACB and ETST component scores. For completeness, four of the cells of the matrix contain r's obtained in the original research defining ASVAB content. Those r's were obtained from 1966 administration of the then operational ETST and the then operational ACB to a full range mobilization sample of size 1200.

Table 3 presents results of administration of the experimental Arithmetic Skill test, showing correlation of this test with the operational clerical and quantitative ability measures in the ACB.

---

[1] For administrative reasons.

Table 1

Correlation of Four Army Interest Scales with Air Force and Navy Scales

Army Scale (N=580)	Air Force Scale (N=279)		Navy Scale (N=301)	
Clerical Interest Scale (CA)	Clerical	.37	Data Processing Technician	-.04
			Storekeeper	.11
Electronics Interest Scale (CE)	Electronic	.53	Electrician's Mate	.16
	Mechanics	.34	Electronics Technician	.19
	Scientific	.50	Sonar Technician	.17
Mechanical Interest Scale (CM)	Mechanics	.45	Engineman	.46
	Electronic	.33	Boiler Technician	.45
			Equipment Operator	.42
Ground Combat Scale (CC)	Military	.40	Synthetic "Combat"	.47
	Outdoors	.38		

Table 2

Intercorrelations[1] among Army Classification Battery Subtests and Electronics Technician Selection Test Elements, N=590

	ETST TOT	ETST MATH	ETST SCI	ETST ER	MK	SK	EI	AFQT	AR	WK	PA	MC
ETST TOT	ETST TOT	84	82	69	66	62	52	65	56	56	42	55
ETST MATH		ETST MATH	53	35	68	50	36*	59	54*	46	38	40
ETST SCI			ETST SCI	38	48	61	50	58	44	52	39	55
ETST ER				ETST ER	35	33	39	34	30*	32	19	34*
MK					MK	47	33	59	56	47	37	37
SK						SK	47	65	49	61	36	43
EI							EI	35	23	33	29	48
AFQT								AFQT	81	84	57	52
AR									AR	55	34	37
WK										WK	26	43
PA											PA	47
MC												MC

[1] Decimal points omitted from correlation coefficients.

*Asterisk indicates correlation coefficients obtained from original ASVAB research, Bayroff, A. G. and Fuchs, E. F. The Army Services Vocational Aptitude Battery, BESRL Technical Research Report 1161, February 1970.

Table 3

Intercorrelation[1] of Clerical and Quantitative Measures, N=279

	XAS	AD	AR	MK
Experimental Arithmetic Skill (XAS)	XAS	37	51	46
Attention-to-Detail (AD)		AD	27	25
Arithmetic Reasoning (AR)			AR	60
Mathematics Knowledge (MK)				MK

---

[1] Decimal points omitted.

DISCUSSION

Interest Scales

It is apparent from the magnitude of the correlation coefficients in Table 1 that virtually no scales of the NVII or VOICE are good direct substitutes for comparably-named Army interest scales, even if adjustments were made for range restriction. The possible exceptions may be the VOICE Electronic and Scientific Scales, correlating in the realm of 0.5 with the Army's CE, and NVII scales for Engineman and Boiler Technician ratings which correlate in the realm of 0.45 to 0.50 with the Army CM, and the VOICE Mechanics Scale, which also correlates with CM at that level. No Air Force or Navy scale correlates in any reasonable range with the Army CA Scale, and obviously the Air Force and Navy would not have Ground Combat Scales. The VOICE does have a General Military Scale and an Outdoors Scale, but these only correlate 0.4 and 0.38 respectively with CC. Similarly, a set of 22 NVII items was identified as involving outdoors, athletic, or military content, but this set only correlated 0.47 with CC.

Since the direct substitution of cross-service scales was not indicated, an item analysis approach was attempted. The VOICE was item analyzed against criteria of the four Army scales in order to identify the twenty or so items most highly correlated with each. This approach yielded correlation coefficients of 0.66 with CA, 0.72 with CE, 0.73 with CM, and 0.68 with CC. These r's are considerably higher, but are not cross-validated.

ETST and ACB

In contrast to results of the interest-scale comparison, Table 2 shows some promising areas of interchangeability among ETST and ACB content sets.

Navy and Army Mathematics Tests--The ETST Mathematics (ETST MATH) subtest and the ACB Mathematics Knowledge (MK) subtest both consist of high school algebra items. As may be seen in Table 2, the correlation between these two measures was 0.68. Given restriction of range factors and the small number of items involved (both the MK and ETST MATH are 20-item sets), this is substantial common variance. Alternate forms of the same tests, when so short, often correlate no more highly. It is also interesting to note that each of these subtests bears the identical relationship with AFQT, r=0.59.

Navy and Army Science Tests--The ACB contains no direct counterpart to the General Science subtest of ETST (ETST SCI). The ETST SCI contains physical science items, whereas the ACB Science Knowledge (SK) subtest contains items tapping knowledge of biology and chemistry. Nevertheless, Table 2 shows that ETST SCI correlates 0.61 with ACB SK; again in a restricted sample and with short tests.

Navy and Army Electrical Tests--The ETST Electricity/Radio (ETST ER) subtest is a very difficult set of items for an Army population. Its mean score in the present sample of Army accessions was just at the chance level. It is quite likely that this very low "ceiling" in the Army group was a major contributer to the extremely low correlation of ETST ER with ACB Electronics Information (EI). We note that the r is only 0.39, while the other two ETST measures correlate with ACB measures in the 0.60's. Clearly the two electrical tests are not good substitutes for one another.

Correlations between AFQT and Selected Army and Navy Item Sets--Because the AFQT is considered to be a test of general ability, the respective correlation between AFQT scores and various ETST and ACB scores are of interest. These correlations reflect the influence of the "g"-factor in the item sets; and, as may be seen in Table 2, the ETST total score (ETST TOT) correlates quite highly with AFQT (r=.65) in the restricted sample. Among the parts, the ETST MATH and SCI sets correlated relatively highly with AFQT (0.59 and 0.58), with ETST ER very much more unique (r=0.34). Interestingly, a similar pattern exists among the Army tests, with MK and SK being relatively highly correlated with AFQT, and EI again quite unique. No total score is computed over these ACB items, so there is no Army equivalent to the total ETST score for comparison with AFQT.

Arithmetic Skill as a Clerical Selector

The experimental Arithmetic Skill test had been devised for possible use as a measure of clerical ability. The arithmetic was kept at seemingly the lowest possible difficulty level in attempts to remove the quantitative loading of the set. It is apparent from Table 3, that either the attempt failed, or this approach to measuring clerical ability in samples of Army input is fruitless. We note that despite the utter simplicity of the arithmetic operations, this item set correlated minimally with the current ACB clerical measure, Attention-to-Detail, and most highly with the quantitative measures in the battery, Mathematics Knowledge and Arithmetic Reasoning.

SUMMARY AND CONCLUSIONS

The interest measures of the Air Force and Navy, the Navy Electronics Technician Selection Test, and an experimental Arithmetic Skill test, were each administered to a sample of Army enlistees. Correlations were obtained between Army interest measures and most promising seeming sets from the Air Force and Navy interest measures, between ETST subtests and ACB subtests of comparable content, and between the experimental Arithmetic Skill test and measures of clerical and quantitative ability. The following conclusions seem warranted:

1. There is no Air Force or Navy interest scale which would be a suitable direct substitute for the Army interest measures presently in use.

2. It is probable that Army Classification Battery and Navy ETST needs can be served by a common set of test items covering high school algebra, and by another common set covering science information. Although there is some service science content specificity--Navy for physics content, Army for biology and chemistry--the correlation between the two sets of science items suggests that they may be tapping the same population variance.

3. The ETST Electricity/Radio subtest is too difficult for Army use, and correlates too low the ACB Electronics Information items to be considered an acceptable substitute.

4. The prospect of utilizing a speed test involving simple arithmetic operations as an Army clerical selector is strongly contra-indicated.

HIGH SCHOOL USE OF ASVAB

by

HARRY D. WILFONG

As many of you are aware, the Armed Services Vocational Aptitude Battery (ASVAB) serves somewhat of a schizophrenic existence. At the present time - and in its current format - in addition to use as a production battery within the Air Force and the Marine Corps - the ASVAB also is used as a counseling instrument in some 16 thousand of our nation's high schools. During school year 1974-75, the battery was administered to over 1.3 million students in grades 9 through 12, with the major proportion being tested at grades 10 and 12. Moreover, appli-cation of the battery for the purposes of vocational counseling has been sanctioned at the State Department of Education level in 17 states. Motives for indorsement range from the fact that the program is offered free of charge to use of the test as a means of assessing salable skills available within the state manpower market.

In an article published in the July 1974 issue of the American Psychologist, (Carver, 1974) describes a dichotomy in application between "Psychometric" and Edumetric" oriented tests. Although Edumetrics is little more than an idea, Carver's point that the Edumetric dimension of tests has been sorely neglected by psychologists involved in testing is well taken, it appears that psychologists have been giving away sound psychometric tests to education for many years when education really needs sound edumetric tests. Generalizing to ASVAB, we have found the battery to be used for both purposes, and in an astronishing variety of applications by civilian educators: It is used as an achievement measure; as an aptitude estimate; as a means of facilitating student career awareness; as a group assessment device; and in various and sundry - sometimes unimaginable - planned and unplanned capacities. To satisfy these varied needs, we have had to quickly devise modifications to tailor the high school testing program to common problems expressed by the counseling community.

Let me cite a few examples of perceived counselor needs, gleaned from personal contact with over 2500 educators and counselors, in some 20 states, over the past 24 months.

In early 1973, an initial area of concern involved the normative base used by the Armed Forces Vocational Testing Group (AFVTG) to report individual scores to high school students throughout the country. The initial "normative" base for ASVAB, built around a sample of 3050 prospective enlistees processed throughout AFEES (Bayroff & Fuchs, 1970), which was, in turn, anchored to the World War II mobilization population, simply did not satisfy a computer normative reporting system proper for use in our high schools. The mobilization population norms did not include females, data were accumulated during a draft influence and had limited application - if any at all - to students tested below the 12th grade level. Although the DOD directive governing the high school testing program (DOD Instruction 1304.12) states that the test may be administered to underclassmen only if seniors are also tested, we must recognize that it is a pragmatic concern of the high school counselor that test results be available early enough to aid the student in his/jer career counseling life. Consequently, it is significant to note that ASVAB was administered to some 329 thousand 10th graders and below during school year 1974-75. To be realistic, we simply would not have a salable or acceptable product if ASVAB were to be offered only to 12th graders.

In recognition of the needs of high school counselors, the ASVAB computer report to the student was modified, to report test results by national percentile, specific grade percentile, and grade and sex percentile. These data are based upon the sample of 771 thousand students actually examined on ASVAB-2 during school year 1973-74. Although considered to be a substantial improvement over the previous reporting procedure, more work is needed to develop an adequate body of ASVAB normative reference information. The "reference data" which we currently report to states, high schools and students, are decidedly based upon large subject samples, but the subjects utilized were incidentally available and obviously constitute a biased non-representative sample of the general population of high school students. They do not even reflect the more specific population of students in secondary or post-secondary vocational schools. Cooperation of school corporations throughout the country is being solicited to further this end for ASVAB-5 norming, on some 40 thousand students.

Another problem cited by civilian educators is a need for data which permits the identification of relationships between specific ASVAB profiles and evidence of orientation toward, and success in, the broad spectrum of occupational areas and levels available within the Armed Services. More importantly is a requirement that such profiles be related more comprehensively to areas and levelsof appropriate civilian occupations. These objectives are viewed to be most worthwhile and, if achieved, an invaluable resource for counseling students exploring and making decisions regarding education and occupation and also to be a definite asset to personnel responsible for the selection and placement of individuals in both the Armed Services and in civilian occupations. Research is planned across the services and in the civilian sector to accumulate such data.

If any of you have actually reviewed Volume 2 of our current ASVAB Counselor's Manual - which supposedly delineates relationships between ASVAB composites and specific service occupations - it becomes quite obvious that the document is of heuristic value only. Quite frankly, we have limited data to demonstrate relationships between ASVAB scores and potential successful occupational placement for a wide range of jobs. As a career education or vocational counseling guide, our existing data base is skimpy, especially where application is attempted at the 9th or 10th grade levels. As a start toward translating the varied service validation studies into a usable, interpretive vocational guide for high school counselors, we - at AFVTG - have programmed space in the new DOD Military/Civilian Occupational Source Book for the description of aptitude profiles associated with specific occupational levels within broad vocational categories. This book is intended to draw together, in a single document, all occupational information common to service jobs - related to their civilian counterparts - along the dimensions of duties, qualifications, job codes, training requirements and life styles. Once published, the DOD source book will supplant existing service specific occupational descriptive brochures distributed to the educational community. Common service job descriptions are cross-referenced with both the Dictionary of Occupational Titles coding system and the US Office of Education's Career Clustering procedure. One of the goals of the document is to interject information concerning service jobs into occupational issuances of other federal agencies, a source of data which has - up to this point in time - been for all intents and purposes - nonexistent. Unfortunately, there is one portion of the source book that we will have to leave vacant, not complete for publication by September of 1975; the part of each job sheet which relates occupational performance to an established profile of aptitude test performance on ASVAB. It is our hope that emergence of the common battery will alleviate many of these problems in the future.

Perhaps if I were to briefly summarize exactly how ASVAB is actually being incorporated into various high school testing programs across the country, we can come closer to what ASVAB really means to the non-DOD user.

Some ASVAB users have opted for the battery solely because it is administered and scored free of charge. It is a recognized fact of educational life that austerity cuts most often are first levied at the so called fringe areas such as guidance and counseling. These users would probably accept ASVAB regardless of its interpretative value or supportable validation base since the program represents "soft" money to the local school district.

You are probably all aware of a renewed emphasis in the arena of career education. We used to call it vocational training, industrial arts or - simply - shop. My own observation is that this emphasis is sincerely motivated, is affecting the basic underpinnings of our national educational system, and will have a significant long-range impact upon both our public school systems and the military services. Under this extremely comprehensive umbrella of career education, the development and public availability of ASVAB was quite timely and fortuitous. Without this renewed interest in vocational instruction, a test battery built within the vitrio of service technical training and classification systems would not have attracted 1.3 million students and 16 thousand school counselors. Under this rubric of career education, our joint battery is being applied in a variety of fashions.

In most school systems, as a function of a grossly overloaded counselor-to-student ratio, students are left to make their own interpretations of test results as best they can. Volume 2 of the ASVAB Counselor's Manual - grouping sample service jobs by the five aptitude clusters - is of some help, but is not generally available to the individual student. Our plans for ASVAB-5 are geared toward a stronger emphasis for student self-interpretation of test results.

Certain other school systems/districts/states, have decided to evaluate student potential - and progress - as related only to regional specific norms. The latter is provided on an as needed basis by AFVTG, and accumulated over successive years of testing. We strongly welcome such programs, for it is only through thoughtful application and research by our users in the educational sector that true applicability of the high school testing program can be empirically established.

We have frequently encountered school systems where standard aptitude tests are applied to serve the various "legal" requirements of accountability. Accountability has almost as many definitions as there are states: Accountability for the use of hours in the school day - where all hours expended in non-teaching activities (including test administration) - must be approved through the State Department of Education. Accountability, where the individual teacher is judged (i.e., evaluated) with regard to his/her capability, solely based upon the gain in student performance over a standard time frame, as measured through change indices reflected through aptitude test scores. We suspect that ASVAB is used in this manner thru test-retest situations in various sections of the country.

We have encountered school systems - in fact, entire states - where the DOD High School Testing Program is used as a supplement to an extremely sophisticated application of accountability, wherein "salable" skills, as identified through aptitude test performance - is intercorrelated - through computer matching - with available job opportunities quantified as resident within the state labor market. In these type programs, our subtests of shop, automotive, electronic and clerical information are applied as specific, quantifiable measures of these so called "salable skills." It must be noted that such an application is made in the absence of verifiable validity to appropriate civilian skill areas.

Accountability, and the ASVAB, is also applied in several states where the state department of education is interested in evaluating the relative performance of competitive schools and school districts - over time - against state-wide standards of administrative efficiency. These uses have nothing whatsoever to do with individual counseling needs, and serve to aid state-wide needs such as curriculum redefinition, distribution of state allocated funds, and - quite frankly - political/administrative requirements (which I lump together in the same category).

A most interesting commentary upon the various applications of the DOD High School Testing Program, as summarized before, is that extremely few approaches are geared toward the individual student and - in turn -

interpretation of student aptitude test results into a viable vocational career path.

From the examples cited, I think you can easily see that ASVAB, as used by the civilian educational community, means many things, serves a variety of purposes and is used in many ways that were probably never imagined by the initial architects of the DOD High School Testing Program.

As service researchers, I ask you to keep these considerations in mind as you build your service related experimental designs, conduct your studies and report your research findings. As we all move closer to a common service screening and classification battery, which will have a civilian counterpart, it behooves all of us to consider the concomitant relationship between in-service research and permissible generalizability into the civilian sector. Several, quite simple, techniques come to mind - all of which can be tied into conventional research methodology:

First, as you perform research on service jobs, it would be most helpful if they could be associated with the appropriate Dictionary of Occupational Titles code; in addition to the traditional MOS, AFSC or NEC code.

Secondly, in validation studies, it would be most useful if significant findings were to be accompanied by expectancy tables or similar "scatter-plot" arrays. I have found extensive use of these type data displays by high school counselors in almost all type guidance situations.

The other suggestions I have, although they may appear overly patronizing, pertain to complete reporting of research findings such as: a comprehensive and full definition of the criterion measure; complete description of any sampling techniques; definition - wherever possible - of the composition of samples used to include sex, ethnic group, age and education.

Such procedures, if adopted by all researchers, will greatly facilitate our job in AFVTG at translating relevant service data into interpretive guides of use to the high school counselor.

As a concluding remark, let me suggest that, as we move closer and closer toward a common service classification battery, it may be appropriate for an inter-disciplinary organization - such as the Military Testing Association - to serve as the forum for drafting the outline of a research system that draws the common service and civilian counseling programs closer together. Such an action is definitely within the charter of the MTA and would decidedly further the aims of the all-volunteer force. I thank you for this opportunity to discuss the various counseling needs as related to the DOD High School Testing Program.

Bayroff, A. G. & Fuchs, E. F. The Armed Services Battery. Technical Research Report 1161, U.S. Army Behavior and Systems Research Laboratory, Arlington, Virginia, February 1970.

Carver, R. P. Two Dimensions of Tests - Psychometric and Edumetric. American Psychologist, 1974. 29. 512-518.

TEC: A Manhattan Project in Educational Technology

Wesley K. Roberts, Warren G. Lawson, and William D. Neal
U.S. Army Combat Arms Training Board

During World War II, the Manhattan District was formed in the U.S. Army Corps of Engineers to administer the predecessor to the U.S. Atomic Energy Commission. The "Manhattan Project" was not only instrumental in developing the most devastating weapon ever employed by one nation against another, it also accomplished prodigious achievements in science and engineering, the impact of which continues today in an increasingly technical society. Similarly, three decades later, the U.S. Army has established another "Manhattan Project." This project was established neither for the development of weaponry nor for the furthering of engineering and science; it is predicated upon the goal of applying all that is known in modern educational methods to the improvement of Army training.

The call for such a "Manhattan Project" in Educational Technology was essentially made by Gagne during an interview with Deterline (1970). Gagne concurred with Deterline that it was time for a "Manhattan Project" in education. He suggested that several projects could be established and that more than one educational philosophy could be used. Gagne saw such an effort as one which would use competent educational researchers who would concentrate on large problems in education. Further, such a project would require complete commitment on the part of both a school system and the community it serves and to have as its purpose the systematic development, tryout and evaluation of new ways of conducting education. Gagne foresaw a necessity in such a project for the provision that the children (students) involved did not suffer and that the schools involved were totally committed to the goal of the system's tryout. The project would have to take priority over all other considerations, including administrative arrangements and laws. The result of such a "Manhattan Project" would be the application of modern educational techniques to selected large scale educational problems.

In reality, Gagne's "Manhattan Project" schools have (in a sense) become United States Army service schools, and his "children" have become soldiers. The Army's "Manhattan Project" in Educational Technology is named TEC, which is the acronym for Training Extension Course, a comprehensive system which provides empirically designed instruction for today's soldiers.

TEC had its impetus into the Army's training system as a result of an identified need to provide prepackaged instruction to train soldiers to perform tasks incumbent in their jobs.

How TEC Began

In August 1971, the Chief of Staff of the Army directed the Commanding General of the Continental Army Command to bring together a board of officers to consider ways of supporting unit commanders in conducting meaningful and exciting training. As a result of this directive, the Board for Dynamic Training was formed at Fort Benning, Georgia. The Board was instructed to prepare a report which would outline innovative training concepts for implementation into the Army. An underlying philosophy of the status of Army training in tactical units was centered on an intuition that "the Vietnam strait jacket" had produced a generation of company commanders, battalion staff officers, and many battalion commanders unprepared by either schooling or experience for service in a peacetime Army.

The Board for Dynamic Training was directed by the Army Chief of Staff to concentrate on combat arms training (Air Defense Artillery, Armor, Field Artillery, and Infantry), where it was perceived the most urgent problems in Army training existed. A major portion of the research completed by the Board focused on the status of individual and small unit training. Key findings of the report (1971) included:

> "Army-wide, training is regarded as only marginally adequate. Individual training is evaluated lowest

> in combat arms units of the Active Army; mission training lowest in Reserve Component units."
>
> "Commanders, especially company commanders, feel they could make training more dynamic if they could devote greater personal attention to its planning and execution, but conflicting administrative and support requirements take priority."

Additionally, it was found that a need existed to provide noncommissioned officers, especially junior sergeants, with training necessary to maintain job and professional proficiency and competence. The need was highlighted in the lack of perceived training support for the Military Occupational Specialty (MOS) Test which is an important and recurring evaluation which has a major role in determining the future career pattern of noncommissioned officers.

The Board for Dynamic Training recommended its own termination and the establishment of the U.S. Army Combat Arms Training Board, which would serve as the organization performing the catalyst function in carrying out the recommendations made in the Board's final report.

Recognizing the necessity to focus on individual job training as a matter of priority, the U.S. Army Combat Arms Training Board initiated in the summer of 1972 a program to provide individualized, performance-oriented training materials for individuals occupying the following Military Occupational Specialties (MOS):

11B	Light Weapons Infantryman
11C	Infantry Indirect Fire Crewman
11D	Armor Reconnaissance Specialist
11E	Armor Crewman
13A	Field Artillery Basic
13E	Field Artillery Cannon Operation/Fire Direction
16P	Chaparral Crewman
16R	Vulcan Crewman

To meet the challenge in overcoming the job performance deficit in these eight Military Occupational Specialties, a small group of officers at the U.S. Army Combat Arms Training Board, using the authority granted them in Army Regulation 10-2, began what was to become the largest single effort ever undertaken in applying a modern educational philosophy to conquer identified learning needs.

In the Army's "Manhattan Project" for Educational Technology practically all restraints have been removed; resources and personnel have been totally dedicated to TEC lesson development, production and distribution. Army service schools and major unit commands have committed themselves to the TEC program and the chain-of-command support is prevalent throughout the Army from the Chief of Staff to virtually every company commander to insure that TEC is designed and implemented as a viable solution to identified deficiencies in job performance.

The first step in the program was to conduct a job and task analysis of the eight MOS. The Human Resources Research Organization (HumRRO) was contracted to systems engineer these eight combat arms MOS. This entailed the development of task inventories and job task data for 93 duty positions. The task inventories were subjected to a validation by incumbents serving in the MOS/duty positions and by various levels of job supervisors.

This effort resulted in the development of a baseline of tasks critical for job performance in both peacetime and combat environments for one-seventh of the Army's total strength.

To provide training for these "critical" tasks the U.S. Army Combat Arms Training Board and four U.S. Army Training and Doctrine Command (TRADOC) schools (the U.S. Army Air Defense Artillery School, the U.S. Army Armor School, the U.S. Army Field Artillery School, and the U.S. Army Infantry School) decided to provide prepackaged self-paced instruction to the soldiers in the force--that is "extension training" or Training Extension Courses. TEC lessons were to be the principal method for filling the educational gap that was found to exist between what was determined as necessary job performance and the current level of job performance in the eight selected Military Occupational Specialties.

Early Training Support Required in TEC

In 1972 the know-how of applying modern educational technologies to military subjects was blatantly absent in the Army. Therefore, the U.S. Army Combat Arms Training Board contracted Deterline Associates, Inc., to conduct workshops for Army personnel who would be involved in the TEC project. These two-week workshops, one in San Francisco and one in Washington, D.C., were designed to teach attendees the application of modern instructional design to lesson development. The educational model taught in the Deterline workshops is known as CISTRAIN, the acronym for Coordinated Instructional Systems Training.

The CISTRAIN methodology was developed by Deterline and Lenn (1972) for the California Community College System and was deemed appropriate for application to Army training development.

A unique characteristic of CISTRAIN is the order of lesson development. The sequence used to develop the lesson is the opposite of that in which the learner encounters the material. Shown below in Figure 1 is the order in which the CISTRAIN model proceduralizes lesson development contrasted against the order in which the learner goes through a lesson.

Order in Which the Learner Proceeds Through a "CISTRAIN Model" Lesson	Order in Which a "CISTRAIN Model" Lesson is Developed
Introduction	Self-Evaluation Items
Descriptive Objectives	Descriptive Objectives
Study Resources	Discussions of Self-Evaluation Items
Practice Exercise Items	Practice Exercise Items
Discussions of Practice Exercise Items	Discussions of Practice Exercise Items
Self-Evaluation Items	Select Study Resources
Discussions of Self-Evaluation Items	Introduction

Figure 1. A contrast of the order in which the learner proceeds through a "CISTRAIN Model" lesson and the order in which a "CISTRAIN Model" lesson is developed.

Personnel attending the Deterline workshops were provided instructional materials which they used to institutionalize the CISTRAIN model at their respective schools for use in the TEC project. The schools were then equipped to replicate the Deterline Workshop on a recurring basis. Modifications in the CISTRAIN technique have since been made at most of the TRADOC schools. This has been allowable insofar as all of the essential parts are included in TEC lesson development.

Follow-Up Training in the TEC Project

A careful study of the CISTRAIN model reveals that it assumes the writing of self-evaluation items based upon given criteria. In order to establish what would constitute the criteria for preparing test items it was necessary to identify specific job tasks for training. This philosophy was in keeping with the procedures given in TRADOC Regulation 350-100-1, the Systems Engineering of Training. Because this regulation fails to provide "how to" information for job/task analysis, the U.S. Army Combat Arms Training Board has conducted several workshops to train TRADOC service school personnel to perform this function for the TEC project. These workshops have normally lasted five days and have consisted of both individual and group instruction. The goal of these workshops is to teach methods that are to be used in job/task analysis which results in task identification, task validation, selection of tasks for training, and the preparation of task analysis worksheets from which self-evaluation test items may be prepared.

A second type of five-day workshop was also found necessary prior to the preparation of the first TEC lessons. These "Lesson Development" workshops taught the methods for writing descriptive (training) objectives, performance tests (which were added beyond the provisions of the CISTRAIN model), self-evaluation items, and how to aggregate instructional material resources such as text material, diagrams, photographs, etc., to be used [illegible] the development of TEC lessons.

Preparation of TEC Lessons

Due to the lack of TRADOC service school personnel resources, a decision was made to have the TEC lessons prepared under contract to civilian educational research and development firms. A similar decision was made regarding the mass reproduction of TEC lessons for distribution.

The generic contracting model which has served the TEC project is included as Annex A.[1]

An explanation of the 28 step lesson development model is found at Annex B. The CISTRAIN Model is used as the lesson development approach in the TEC project. It is important to note at this juncture that considerable interaction occurs between the Army (government) and the contracting firm during the preparation of each TEC lesson. Any modifications that occur in the contracting procedures are completed after insuring that no compromise is made in the quality of instruction in the TEC lessons.

A small number of lessons have been prepared totally within some service schools. These have been restricted mainly due to the lack of personnel skills and resources which can be made available to perform the functions normally contracted in TEC lesson development.

Growth of the TEC Project

The TEC project has essentially grown in four phases. This growth has occurred as a result of a planned expansion based upon; (1) the phased application of resources for the project, (2) the phased development of managerial precedents, and (3) an evaluation of the success of each preceding phase.

TEC I was the initial project developed by the Infantry School. Fifty-six lessons, applicable to Infantry NCOs in MOS 11B40, were produced in a synchronized sound/slide format by Infantry School instructors and Advance Course students and fielded in ten Active and Reserve Component Infantry battalions. The lessons were prepared in a crash effort to provide materials to the field. As a result, they were not derived from an MOS definition effort, were not validated, and were not performance-oriented. Each lesson was, in effect, an illustrated lecture. All of these deficiencies were recognized at the time. Use of the lessons in these units was then measured by a HumRRO (1973) research effort to ascertain their acceptability and to evaluate effectiveness in upgrading MOS scores. This HumRRO research effort pointed out several significant findings:

---

[1] Annexes A and B have not been included in the proceedings. Persons desiring copies of these documents should contact the senior author.

- The results demonstrated strong support for the concept when employed in a favorable command climate.

- Those units which required soldiers to use the TEC lessons or incorporated the lessons into their training programs scored higher on the MOS test than did the units which failed to take advantage of the training materials.

- Use of the TEC I training materials proved effective in reducing the effect of verbal English ability on MOS test performance.

In addition, the HumRRO report recommended improvements for making the TEC concept more effective in future applications. The report concluded that:

- TEC materials should be systems engineered around specific performance objectives to accomplish a job relevant purpose and pretested to insure these objectives can be met by soldiers who take the lesson.

- Guidebooks should be developed for trainers to assist them in integrating the system in their unit's training.

As information on the TEC I program reached the other nonevaluation units in the field, demand for the system was extensive.

The results of the HumRRO research report, analysis of the costs associated with meeting continuing demand for the TEC I lessons, and advisements from some of the leading educators in the country prompted the development of three conclusions from the TEC I experiment.

- The concept of exporting service schools prepared lessons on individual skills is viable....schools have the ability to produce these lessons and commanders in the field need and desire them.

- Materials prepared as exportable packages for field use must concentrate on the critical and important tasks that a man is required to accomplish in performing his MOS responsibilities...MOS must therefore be more definitely defined than ever before.

.Future lessons exported to the field must be built on the precept that soldiers learn most from doing...lessons must therefore utilize the latest techniques of instructional technology in developing individually paced, soldier tested, performance-oriented materials.

Using these conclusions as the base, decisions were made to find a more effective yet less costly method of expanding the TEC project. The resulting program was a validated instructional program referred to as TEC II.

TEC II was directly related to the findings on the TEC I experiment. This entire program has incorporated the lessons learned from the USACATB and HumRRO research efforts and is geared to satisfy demands from the field to expand the development of prepackaged, self-paced instructional materials at the service schools.

The TEC II expansion included the four combat arms schools: Infantry, Armor, Field Artillery, and Air Defense Artillery.

Unlike TEC I, TEC II lessons are based on an extensive systems engineering effort to determine the critical tasks which should be trained. Although the TRADOC service schools have continuously conducted a form of systems engineering under TRADOC Regulation 350-100-1, these procedures were challenged under the TEC development. In the past, the traditional approach at the school has driven the teaching points. In TEC, lesson objectives were derived from task content measured by job surveys. Each MOS has been totally redefined in terms of the tasks that a man must be able to accomplish to survive on the battlefield or accomplish his job requirements. This redefinition of the MOS is a lengthy process taking over a year's time during which a number of MOS definition changes have been isolated, impacting not only on TEC II, but also into the schools' internal operation and doctrine development.

Perhaps the most significant change in TEC development as compared to materials produced in the past, involves lesson validation. Each TEC II lesson has been validated before distribution to the field by tryouts on soldiers in the MOS for which it was designed. If the lesson does not meet the designated validation criteria, it is refined until it does. This validation process is one of the most important differences between TEC I and TEC II.

Following introduction of TEC I and II concepts into the field, combat arms commanders indicated a need for training assistance in a number of high skill level jobs which are found in small numbers in every unit. A survey was conducted to ascertain the specific, yet most critical, areas of need. Survey results prompted the further expansion of the TEC project in the summer of 1973 to each of five critical duty, low-density positions found in combat arms battalions--the company clerk, radio teletypewriter operator, TAMMS clerk, PLL clerk, and generator operator. This expansion was designated TEC III and encompassed the proponent service schools of each of the critical duty jobs; i.e., the Administration, Signal, Ordnance, Quartermaster, and Engineer Schools. Under TEC III, these service schools, following the TEC II model, went about the system engineering effort designed to specifically define the job requirements in the field by survey, and began training their staff.

Initial distribution of lessons produced under the TEC II project reached the field in June 1974. Upon fielding of these lessons, commanders of combat support and combat service support units and schools actively sought participation in the project. Recognizing the heavy front load effort required in developing TEC lessons, TRADOC initiated actions to expand the program further in FY 76 and the outyears. As a result nine additional schools have been added to the TEC program, and designated as TEC IV schools. Each of these TEC IV schools is currently conducting detailed analysis to determine subjects which should be developed in a TEC format and which tasks are critical for those subjects which have been selected for lesson development. As with all of the previous expansions of the program, lessons learned in TEC I-III are being structured into the TEC IV development process.

Recognizing that TEC is not uniformly applicable to all environments, that field units are not the sole user of the system, and that TEC developments must be orchestrated with other training programs, additional guidance has been prescribed for service schools to utilize in selecting areas of TEC lesson development.

.Jobs or MOS that pose significant training problems in the field (Active Army and Reserve Components) should be a prime criteria for lesson development.

.Lessons selected for development should be consistent with the EPMS training strategy for that MOS.

.Lessons should be not only viewed from the standpoint of providing school expertise to the field, but also from the possible standpoint of providing remedial training and in some cases replacing institutional training at Army training centers and schools.

Development should concentrate on training to improve weapons systems effectiveness. This criterion emphasizes the knowledge that the best weapons in the world will not be effective unless they are operated by well trained soldiers.

Implementation of TEC

Implementing TEC into the total Army (including Reserve Components) has been a formidable task. Notwithstanding the complexities of introducing the changes associated with this new system, the sheer number of units involved (over 1500) prompt the requirement for detailed implementation planning. As a result, detailed procedures were designed to introduce TEC Army-wide. Many aspects of the implementation plan are ongoing and will continue into the near future.

Effect on the Training Problem

The results of TEC on the Army's training problems has yet to be fully realized and evaluated. Initial evaluation data shows a trend in narrowing the desired and actual job performance gap when TEC lessons are used in training. A standard of TEC lesson use in the conduct of Army training is continuing to emerge. The research completed to date by the Army Research Institute and Research for Better Schools, Inc., provides substantial data supporting the Army's use of TEC as a training technology. The most recent cumulative results from large group lesson validation trials (Figure 2) show significant improvement on test scores after only one lesson trial.

Type Lesson	Number Lessons	Pretest	Posttest	Number of Soldiers Tested
Common	68	32.4%	89.1%	1722
Armor	48	39.5%	86.6%	1098

Type Lesson	Number Lessons	Pretest	Posttest	Number of Soldiers Tested
Field Artillery	53	40.6%	92.6%	1023
Air Defense	53	43.7%	93.6%	1083
Infantry	27	45.2%	87.0%	690
TOTAL	249	39.1%	91.1%	5616

Figure 2. Cumulative validation data from large group trials of TEC lessons based on one exposure to the instructional materials.

To date optimal methods for TEC use can only be theorized based upon the embryonic state of its implementation into the Army's training system. At this time, the United States Army Combat Arms Training Board is sponsoring research through ARI to develop cost efficient TEC implementation and utilization models for optimizing TEC utility in field units. Additional research is being proposed to determine how to optimally utilize TEC within the resident training environment.

Continuing Program Evaluation

The enormity, complexity, and novelty of the TEC project within the military environment dictates an ongoing and continuous evaluation program. TEC evaluations vary in detail and scope and usually can be categorized into the following areas.

- Product (lesson) evaluations
- Process evaluations
    - Design
    - Development
    - Production
    - Distribution
    - Implementation
- Utilization and utility evaluations
- Evaluation of management procedures (Planning, organizing and controlling)
- Evaluation of staff development and staffing requirements.

Many of these evaluations are being conducted by the U.S. Army Research Institute. Others are being conducted by USACATB personnel and professional educators and educational psychologists employed under the Army Research Office's Scientific Services Program.

Each year since 1972, five to ten percent of the total TEC budget has been spent for these evaluations. This trend will apparently continue for at least the next three years, which points out a significant aspect of the application of modern educational technology to new environments--program evaluation requirements, and the costs of meeting these requirements, tend to expand rapidly and consume a significant amount of total program resources. This is not unexpected, but often overlooked by program managers.

Epilogue

The Army's "Manhattan Project" in Educational Technology remains incomplete. TEC's momentum has yet to reach a point to where it is self sustaining. In the course of designing and developing the TEC system the Army has accumulated a wealth of experience relative to the utilization of systematic instructional design, the management of validated instructional product design and development, and the implementation of a totally new and comprehensive instructional technique addressing individual and collective training requirements for the modern soldier.

A myriad of competent educational researchers from the military and civilian communities, to include Deterline and Gagne, from whom came this whole notion of a "Manhattan Project" in Educational Technology, have contributed to the TEC project.

This educational research and development represents the largest single effort to substantiate earlier findings on the merits of systematic instructional design and development undertaken to date by any single institution. While the system is now in its early stages of utilization by soldiers, its applications are expanding as the quality of the product continues to be verified.

REFERENCES

Deterline, William A., What Students Learn - and How: An Interview With Dr. Robert M. Gagne', Sound Education Reports, Volume 1, Special Report No. 2: 1970.

Deterline, William A., and Lenn, Peter D., "Coordinated Instructional Systems", Palo Alto, California: Sound Education, Inc., 1972.

HumRRO Technical Report 74-16, "Test and Evaluation of the Training Extention Course, June 1974.

Report of the Board for Dynamic Training, VOL I - Executive Summary. Fort Benning, Georgia: US Army Field Printing Plant, 1971.

A Comparison of TEC and Conventional Training by Use of Performance Tests

by

John J. Kessler, Claramae S. Knerr
and Ronald G. Downey

U. S. Army Research Institute for the Behavioral and Social Sciences

Introduction

The Army's Training Extension Course (TEC) System contains audio-visual lessons designed to train and improve the soldier's job skills. Each lesson is based on job requirements determined through systems engineering of MOS. Currently programmed are about 600 lessons, directed at combat arms and common subject matter, which are expected to be in the field by the end of 1976. The Army Research Institute conducted a field experiment which used specially-designed performance tests to compare the effectiveness of TEC training with training carried on by conventional instruction (CI) and to compare both against an existing baseline (BL) level of performance.

Procedures

The samples of soldiers selected for inclusion in the experiment were drawn from the Active Army and the National Guard. They were chosen on the basis that the topics of the training were relevant to their jobs and that they were representative of units which future TEC training materials would serve. Five topics were trained during the course of this research: Light Antitank Weapon (LAW), 81mm Mortar Fire Direction Computer (FDC), M16A1 Rifle, Hand Grenades, and Surveyed Firing Charts. The first four were covered in both the Active Army and the National Guard; Surveyed Firing Charts was covered in the Active Army only.

At each site, the treatment groups were formed by dividing approximately 60 enlisted men into three groups with roughly comparable distributions of pay grades and then randomly assigning each group to either TEC training, conventional training, or the baseline group which received no training on the topic under investigation. The TEC group was taken to a room where each man had his own Beseler Cue/See machine and supply of audio/visual lessons. The Conventional Instruction group was taken to a room where an instructor designated by the unit carried out training which he had planned and prepared. Other than a list of training objectives, he had no specific directions. Individuals in the baseline group returned to their normal duties or received some form of cross training.

Figure 1 shows the flow diagram of experimental procedures. In the Active Army units, these procedures took place over two weeks. In the National Guard units, the procedures were confined to one weekend. The pre- and posttests indicated refer to paper-and-pencil tests which accompanied each TEC lesson and are part of TEC's instructional technique. In order to equalize possible advantages that the TEC group might obtain from merely taking these tests, the Conventional Instruction group and the Baseline group also received these tests. They did so in a combined booklet form rather than after each segment of instruction. Additional efforts were made to control and equalize other factors which might bias the training results. The most important of these was to provide unit instructors with training objectives which were edited versions of those employed by TEC. Time duration of Conventional Instruction was limited approximately to the TEC time period, and approximately equal on-site physical facilities and materials were made available to both TEC and Conventional Instruction groups. The TEC groups were confined to using only the equipment called for in the TEC instructions.

Performance Tests

The criteria of training effectiveness and the measures of primary interest in this study were criterion-referenced performance tests. Using the training objectives as bases, these tests were constructed from proven existing tests and from systems engineered performance objectives. The latter, because they were the most recent and complete performance objectives available at the time of the experiment, served both as a source of updating information for the existing tests and as focal points for additional new test material. In the case of some training objectives, neither proven performance tests nor corresponding validated performance objectives were available. As a result, the scope of the performance tests ultimately used in the field experiments tended to be narrower than the scope spelled out by the training objectives. The format used in all the performance tests is the one used in the Soldier's Manual Army Testing (SMART) books and is described in Guidelines for the Conduct of Performance Oriented Training (TRADOC Pamphlet 600-11). In this format the test situation, conditions, necessary equipment and standards, are specified and the performance requirements are described. Each item is dichotomously scored, either "go" or "no go", and a soldier's final score equals his total number of "go" items.

Administration of the performance tests was carried out by soldiers from the 100th Division (Training), U. S. Army Reserve, Test Section and the U. S. Army Field Artillery School. Their rehearsals and practice with the tests were invaluable in anticipating potential difficulties. A further benefit of this kind of preparation was the development of consistent scoring policies by the test administration teams. Since performance testing can be very complex, expensive, and time consuming, trained and reliable test administrators such as these are absolute essentials.

Results

The results of this experiment are shown in Figures 2 and 3. The numbers refer to the mean per cent of correct answers for each group on each performance test. Active Army results (Figure 2) and National Guard results (Figure 3) are displayed separately because of differences in time allotted to complete the experimental procedures and different on-site conditions. The general pattern was for the TEC group to perform best, followed by the conventionally trained men, and finally the Baseline group. TEC trained groups always scored higher in performance than the Baseline group, whereas the conventionally trained were somewhat more variable in comparison to the Baseline group.

Another result which can be observed in these figures is that where the nature of the performance test was highly cognitive, i.e., Mortar FDC, Surveyed Firing Charts, and Hand Grenades, the TEC trained groups achieved their widest margins over the conventionally trained groups. Figure 4 shows an example of the cognitive or "mental" quality of the hand grenade performance test in that it involves selection and identification of appropriate grenades for particular combat situations. By contrast, the Light Antitank Weapon (LAW) and the M16A1 Rifle performance tests involved primarily psychomotor tasks. An example of this kind of task is shown in Figure 5, a portion of the LAW performance test which requires the soldier to open up a LAW and position it for firing. Differences between TEC and Conventional Instruction in training effectiveness on the latter type of performance were relatively small.

In addition to the performance test scores, data on other relevant variables and covariates were obtained. The list of all experimental variables is shown in Figure 6. The Pretest and Posttest refer to the written tests mentioned earlier which accompany TEC lessons. No significant differences were found between training groups for any of the samples on the Pretest variable. On the written Posttests, as was expected because they were directly tailored for TEC lessons, the TEC trained groups scores significantly higher. General mental ability, as measured by GT scores, was obtained from the soldiers' records and was used in a covariate analysis of the performance scores. In only two cases, however, were significant differences among training groups found on GT: Mortar FDC in the Active Army and M16A1 Rifle in the National Guard. Adjustment of performance score means for GT differences in these groups produced no appreciable changes in the basic results. Inspection of the frequency distributions of the background variables (Primary MOS, Pay Grade, Time in Service, and Previous Training) also showed nothing which would suggest that the composition of the samples might bias the results.

During the analysis of data relating performance test scores and GT scores, an interesting result was observed. As shown in Figure 7, correlation coefficients were calculated between these two variables for the training groups within the Active Army. A clear pattern can be seen: for the TEC trained groups these coefficients are all very small, ranging from -.13 to +.06, whereas for the Conventional and Baseline groups they are generally much higher and positive, ranging from +.11 to +.71. Recalling that the TEC performance score means were nearly always above those of the other two groups, this pattern suggests the TEC training was effective for men of both low and high general mental ability while the effectiveness of conventional instruction tended to be dependent on GT. In the National Guard this pattern was not observed, possibly because few men were at the lower GT levels or because the time intervals between the experimental procedures were different from those in the Active Army.

Summary and Conclusions

To summarize, the primary purpose of the field experiment was to compare the training effectiveness of the Training Extension Course's automated audio-visual methodology with that of conventional instruction and to compare both against the existing baseline level. To do this, five topics were taught at Active Army sites and four at National Guard sites. Training effectiveness was assessed by means of carefully designed and administered performance tests. These tests, because they were composed from proven performance tests already in existence and tied to systems engineered performance objectives, provided measures of what is very close to the final objective of training, job performance.

It was concluded that the TEC trained groups' performance was consistently better than or equal to the conventionally trained groups' performance. When both were compared against baseline levels of performance, the TEC groups' margin was more uniformly maintained across the topics. Although the net overall effect of both training methods was positive, TEC appeared to have an advantage in dealing with material of a predominantly cognitive nature. The results in the Active Army also indicated that the TEC training was both more effective and independent of general mental ability as measured by GT level. More experimentation, however, is needed before this finding can be regarded as conclusive.

In the near future, the Army Research Institute will publish a more comprehensive report of the research just described. A highly detailed and thorough analysis of comparative training costs will be presented in addition to the research on training effectiveness. Other related efforts include a project in which the statistical properties of the performance tests used in the current study are being analyzed for the purpose of identifying more efficient measures of skill mastery. Also underway is a project in which the learning retention of TEC trained individuals is being monitored at points three months and six months after training. Employing both written tests and performance tests, this project is examining the relationships between retention rates of knowledge and performance ability. These research efforts are expected to add to the Army's knowledge of training methods and ways to accurately evaluate their outcomes.

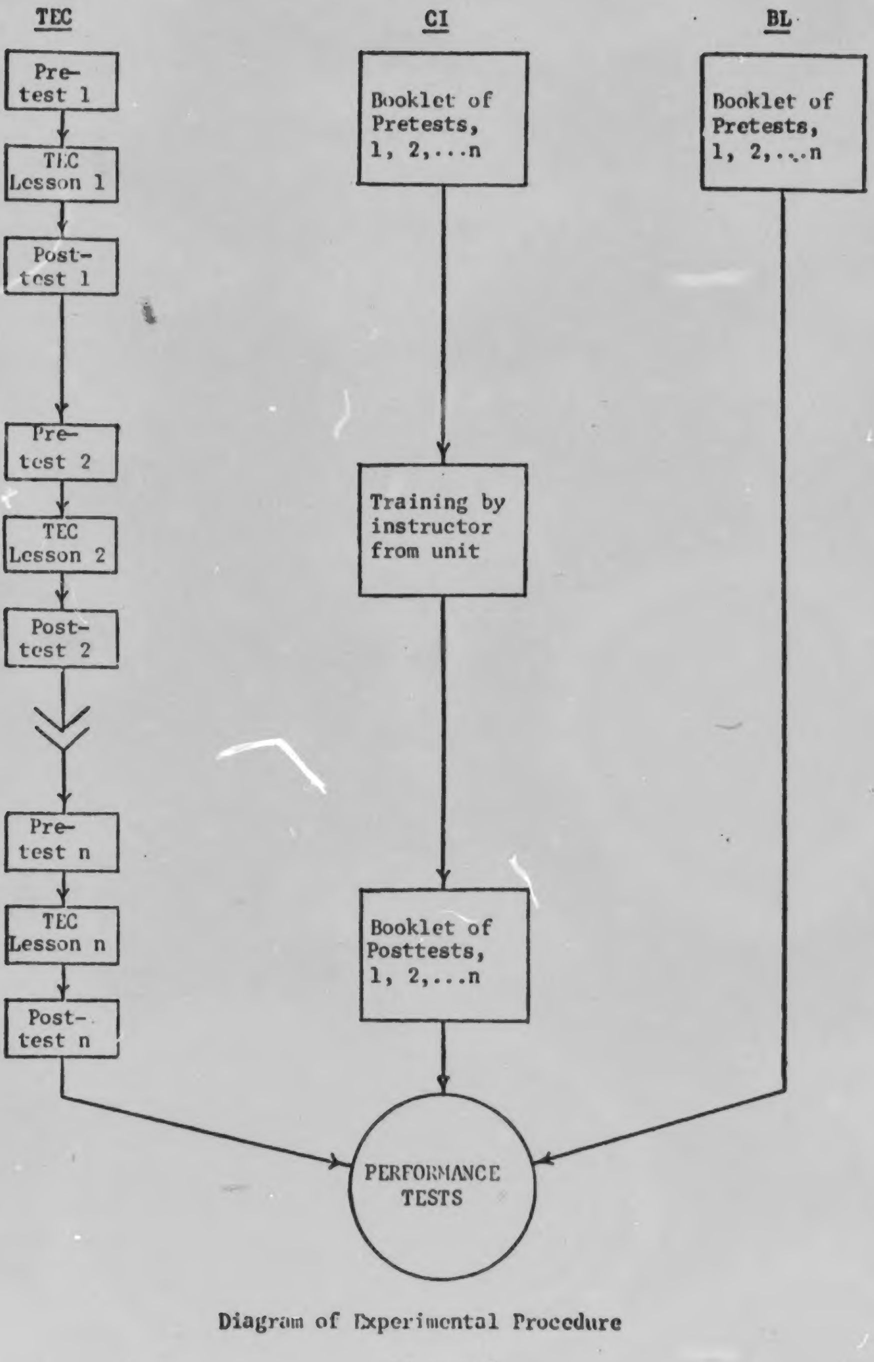

TEC
CI
BL
Pre-test 1
TEC Lesson 1
Post-test 1
Pre-test 2
TEC Lesson 2
Post-test 2
Pre-test n
TEC Lesson n
Post-test n
Booklet of Pretests, 1, 2,...n
Training by instructor from unit
Booklet of Posttests, 1, 2,...n
Booklet of Pretests, 1, 2,...n
PERFORMANCE TESTS

Diagram of Experimental Procedure

Figure 1

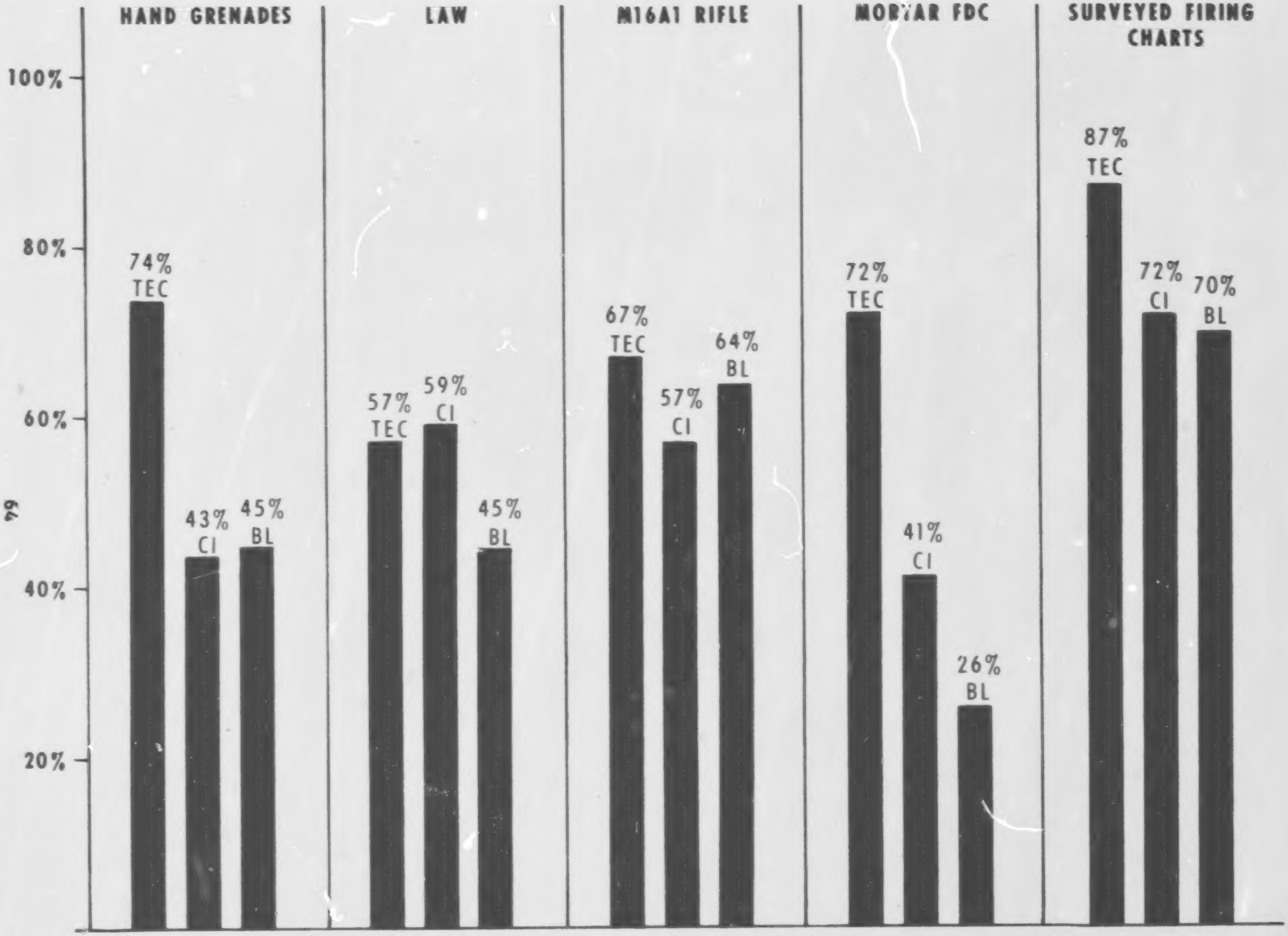


Figure 2. Active Army performance test score means expressed as percent correct by experimental group and lesson category

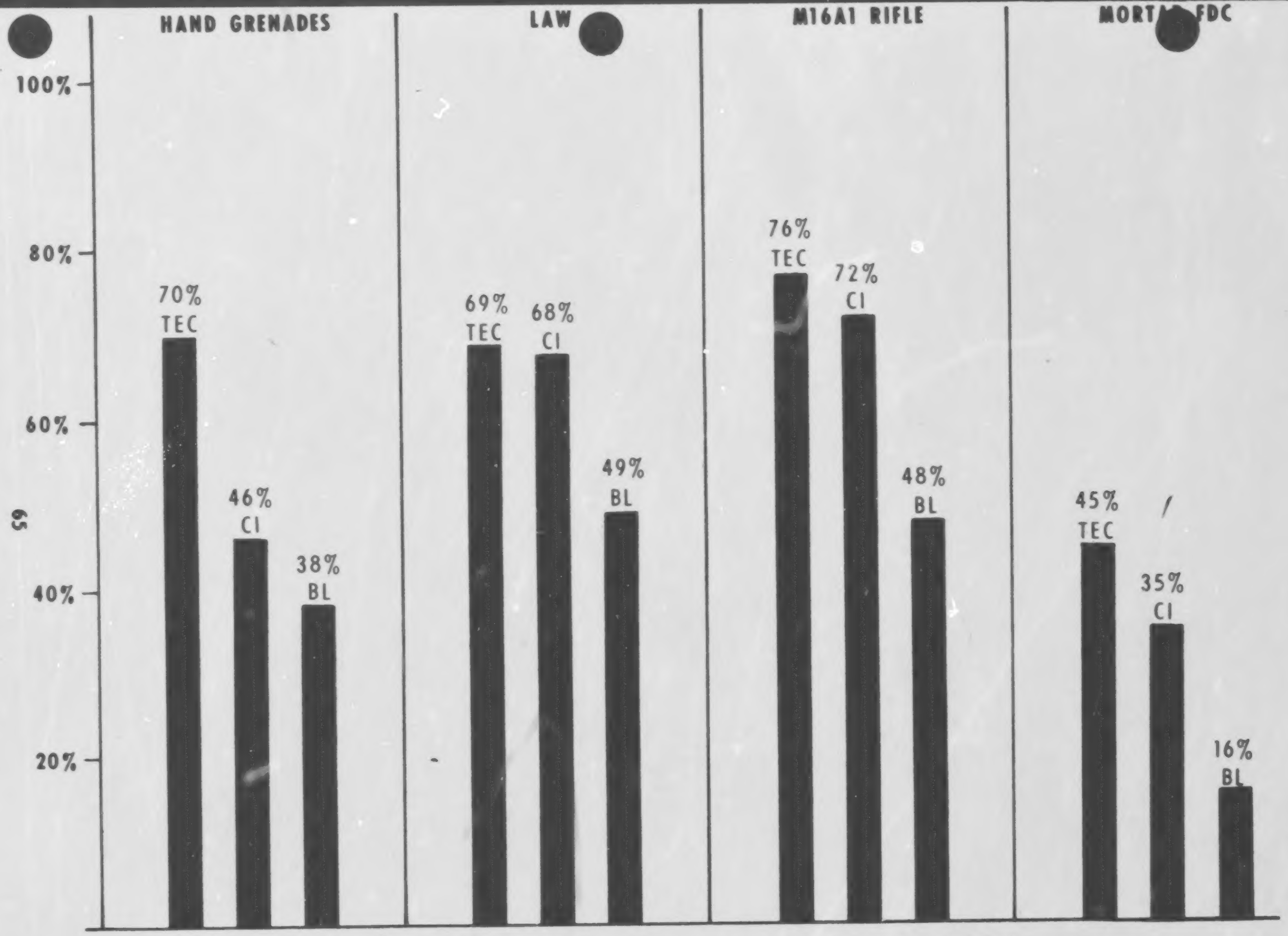


Figure 3. National Guard performance test score means expressed as percent correct by experimental group and lesson category

Necessary Equipment:

Examples of the hand grenades named under the test conditions.

Performance Measures:
The tester will read the situation in quotation marks to the soldier. The soldier will name and point to the correct hand grenade to use for each situation.

	GO	NO GO
A. "You are in a recon patrol that has made contact; you want to break contact and cover your withdrawal. Name and point to the grenade that you would use."		
B. "You are rehearsing the conduct of a raid; aggressor forces will be used. Name and point to the grenade that you would use."		
C. "You are signaling for an aerial medical evacuation mission. Name and point to the grenade that you would use."		
D. "You need to prevent enemy use of a mortar barrel. Name and point to the grenade that you would use to destroy it."		
E. "You are attacking enemy troops located uphill and 25 meters from you. Name and point to the grenade that you would use."		
F. "Civilians are holding a power station; you are attempting to retake that station; you want no casualties and no damage. Name and point to the grenade that you would use."		
G. "You are to destroy an enemy supply point of flamable items. Name and point to the grenade that you would use."		
H. "You are attacking enemy personnel who are in tunnels; you want prisoners. Name and point to the grenade that you would use."		

Figure 4. Sample page from hand grenade performance test.

M72A2 LAW

Performance Test 1: Placing the LAW into Operation

Test Situation
(tester will read to soldier)

"At this station you must place the LAW into operation, demonstrate firing positions and misfire procedures, return the LAW to the carrying configuration, and select targets for the LAW. Your first requirement is to place the LAW into operation. To perform this test you will be using an expended M72A2 LAW. Assume that direction is down range. Commence work."

Test Condition:

Tester will point in direction assumed to be down range. The soldier will place an expended LAW into operation.

Necessary Equipment:

One expended M72A2 LAW.
The launcher will be collapsed, retaining pin in the rear cover, and the sling attached.

Performance Measures:

	GO	NO GO
A. Remove the rear cover retaining pin.		
B. Rotate the rear cover downward.		
C. Remove the sling assembly and front cover.		
D. Extend the launcher until it locks in the extended position by grasping the rear sight cover and pulling sharply to the rear.		
E. Check the backblast area to the rear of the weapon and place the launcher on the shoulder.		
F. Arm it by moving the safety handle to the ARM position (may simulate firing).		

Test Standard:

Performance measures must be performed in sequence. If the LAW does not fully extend, the soldier may try as many times as needed to cock the LAW. The soldier must keep the weapon pointed down range and insure that all parts of the body are clear of both ends of the weapon.

Figure 5. Sample page from LAW performance test.

Variable	Description
Performance Test	Hands-on measure of skills after training; scored by the total number of "go" points.
Pretests and Posttests	Written tests that accompany the TEC lessons; given before and after training and scored by the number of correct answers.
General Technical Aptitude Area	Composite of verbal and arithmetic aptitude from the Army Classification Battery; scores obtained from soldier's records.
Background Information:	
Primary MOS:	11B, 11C, 11D, 11E, 13A/B, 13E, or other
Pay Grade:	coded E1 to E9
Time in Service:	Intervals of 1/2 year up to 4 years, and more than 4 years
Previous Training:	No training on the research topic; BCT only; BCT and AIT; other training on the topic but not BCT or AIT; BCT, AIT, and other training.

Figure 6. Experimental Variables

Lesson Category	Experimental Group		
	TEC	CI	BL
Hand Grenades	-.02	.61*	.24
LAW	.06	.54*	.44*
M16A1	.02	.24	.43*
Mortar FDC	-.13	.27	.71*
Surveyed Firing Charts	.02	.11	.27

* Significantly higher than zero; $p < .05$

Figure 7. Active Army Correlations Between Performance Test and GT Scores, by Lesson Category and Experimental Group

Discriminant Analysis Applied to Sequential Testing

by

David B. Thomas, Ph.D.
University of Iowa

Support for this research was provided by the Graduate College and the University Computer Center at the University of Iowa. The author would like to express his appreciation to the Iowa Testing Programs for their Cooperation.

Discriminant Analysis Applied to Sequential Testing

David B. Thomas, Ph. D.

Computer Assisted Instruction Laboratory
University of Iowa

Individually prescribed instruction programs have imposed an increased reliance on tests as a means of selecting students and routing them through the sometimes complex programs. Instructional prescription models are being developed which depend upon the assessment of various learner characteristics including achievement, personality factors, aptitudes, attitude, and personal interests. Assessment tests are scheduled prior to instruction and may be rescheduled a number of times during the instructional process. Test batteries are also administered to select students for a particular instructional program - mechanics or electronics for example.

The potentially large number of tests that could be administered as part of a training program conducted within an individualized framework could easily impose a burden on resources. Costs of administering, scoring, and interpreting tests and of supporting trainees could reach a substantial amount. Reducing the time devoted to testing may therefore be considered a high priority activity. It is important, however, that the amount of information obtained through a testing program not be reduced significantly.

Testing which occurs within the training sequence is particularly vulnerable to inefficient use of both trainee and instructor time. Students complete units and could be tested at different times, perhaps over different content skills. Scheduling and monitoring of imbedded tests could present a major administrative problem.

Computer-based instruction systems provide one means for monitoring and scheduling the complex routings which may take place within an individualized training system. Test administration may also take place under computer control. The turnaround time required for scoring each test could be eliminated by programming the computer to score the test as it is administered.

A computer-administered test is not necessarily a shorter test, and although a computer-based system may reduce administrative costs somewhat, a means to reduce the testing time from the student perspective is desirable. There have been numerous attempts to shorten testing time: short linear tests composed of the most highly discriminating items from a longer test; branched or tailored tests in which items are presented with a difficulty level selected so as to approximate a dynamically estimated examinee ability level; and sequential tests in which items are administered until a criterion is satisfied. Branched tests are described by Wood (1973) and Weiss (1974). Cleary, Linn and Rock (1968) describe short linear tests and a sequential test based on a procedure developed by Wald (1947). Branched tests require that a large item pool be developed for a test of moderate length. Short linear tests reduce testing time, but at the expense of non-adaptability to student item responses. The sequential test is adaptive to each examinee's response protocol and requires a small item pool. This paper presents the rationale for a specific sequential testing model, describes a validation of the technique, and discusses implications of the procedure for military training programs.

The Sequential Testing Model

The sequential test may be described as one in which an examinee is given a test item, the response to the current item as well as previously administered items is evaluated in a certain way, and the examinee is either assigned to one of a number of classes and testing stopped, or judgment is suspended and testing continues.

One method which may be employed to classify examinees requires the use of a discriminant function (Tatsuoka, 1971). Discriminant analysis is a procedure which determines the coefficients of a linear combination of variables (the discriminant function) which best discriminates between groups of individuals. Typically, discriminant functions are derived employing test scores from a battery of tests. If single items from a test are employed rather than single total test scores from a battery, then functions could be computed to provide classifications after a small number of test items had been administered. If coefficients were computed for linear combinations of all items administered at any given time, such as a linear combination of the first six items, the first seven items, and so on for as many items as are in the test, then each individual could be assigned to a classification using the discriminant functions sequentially.

As the number of variables in a discriminant function must equal or exceed the number of assignment classes, n-g sets of functions would be required for a test composed of n items and used to classify individuals into one of g groups. If these sets of functions are numbered g, g+1, g+2, ...,n, then an individual could be administered g test items, the items could be scored and the scores converted to linear combinations using discriminant function set g. If the assignment were to be into one of four classes, for example, then four values, posterior probabilities of group membership, would emerge from the discriminant functions, each probability corresponding to one of the four classes. The individual could then be assigned to the class indicated by the highest probability of group membership.

The approach described provides a potential assignment following each test item administered, for the technique results in probabilities of group membership every time a linear combination is computed. Thus, one could administer four test items, compute probabilities of group membership, assign the individual to the appropriate class, and stop testing. As an examination of Table 1 indicates, this action would not be desirable. Consider the two individuals whose scores are indicated in Table 1. Person 1 would be assigned to Group 3, as the probability of membership in Group 3 is .90. If the discriminant function had been shown to be accurate in the past, one would probably be willing to accept that classification and terminate testing. Person 2, on the other hand, would also be assigned to Group 3 as the maximum probability of group membership, .55, corresponds to Group 3. Clearly, additional testing would be desired for this person as the probability of group membership in Group 3 is low. Thus, it is desirable to adapt the discriminant function procedure to a sequential testing procedure using the probability of group membership as a guide.

The discriminant function procedure may be adapted to sequential testing by the addition of a termination rule. The varieties of termination rules are endless, but a natural class of termination rules would be related to the probability of group membership. For example, testing could continue until the maximum probability of group membership reached a certain level, say .85. In the example given in Table 1, this termination rule would permit Person 1 to be assigned to Group 3 after four items had been administered and testing would terminate for that person. Person 2 would be administered four items, but judgment would be suspended as the maximum probability of group membership, .55, did not meet or exceed the minimum required for termination of testing. Person 2 would be administered another item, and additional items, until the maximum probability exceeded .85. To assure that an assignment was made and item economy maintained, an additional rule could be incorporated which would terminate testing following some maximum number of items, say 30, if termination had not already occured. Other termination rules could be suggested, but the rule mentioned is easy to use and interpret and can serve as a point of departure for research in this area.

Table 1

Classification of Two Individuals
by Discriminant Analysis

	Class			
Person	1	2	3	4
1	.01	.04	.90	.05
2	.05	.10	.55	.30

Entries are posterior probabilities of group membership.

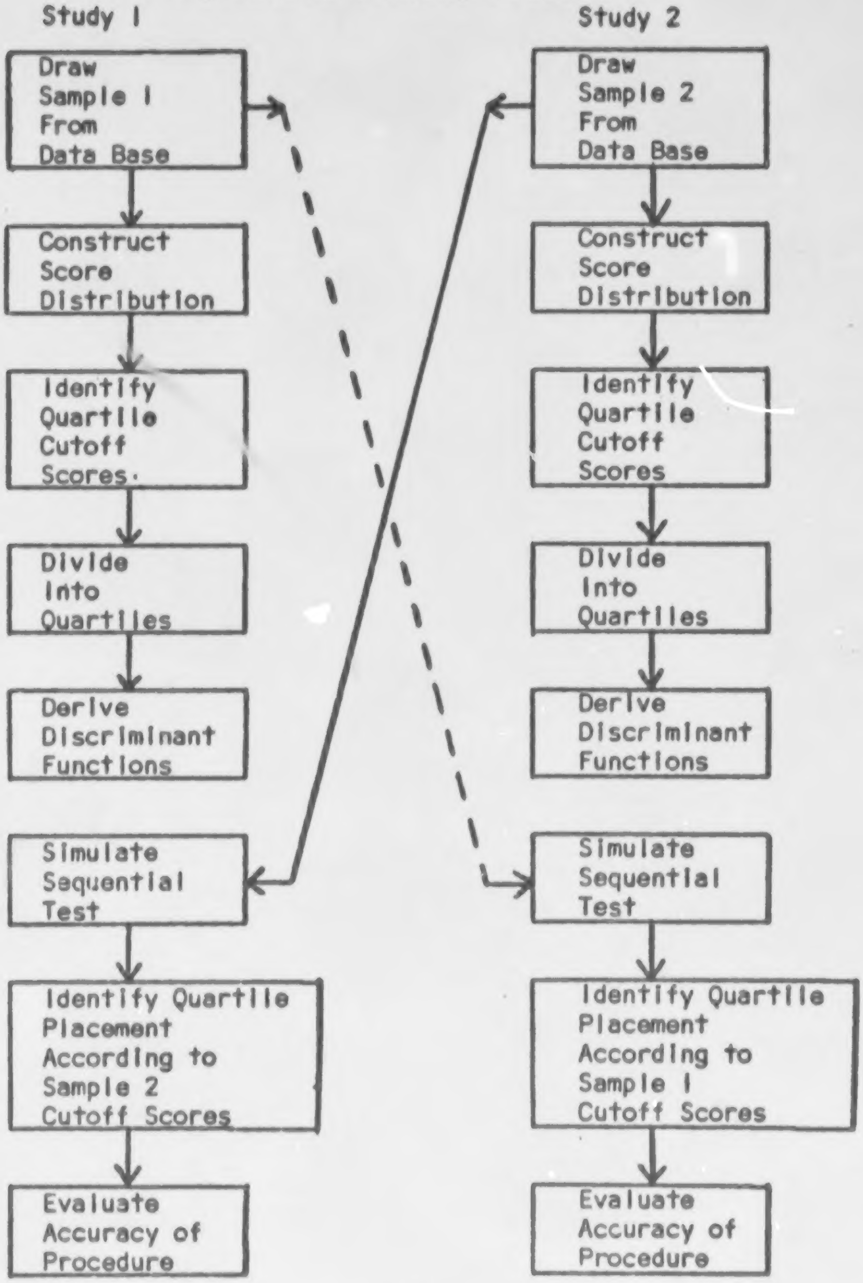
Method

Item response data from the 1971-1972 administration of the Iowa Tests of Basic Skills (Hieronymus and Lindquist, 1974) were employed to simulate the sequential testing procedure. Two samples representing 256 examinees each were drawn from a data base consisting of the response protocols of over 24,000 examinees who had completed Form 6, Level 14 (Grade 8) of the test battery. The vocabulary section consisting of 48 items was used in this investigation.

Two identical studies were conducted, with the second study constructed as a mirror-image of the first. The total number of correct responses was identified for each examinee. Score distribution tables were derived in order to identify

FIGURE I

Sequential Testing Procedure


Study 1
Study 2
Draw Sample 1 From Data Base
Draw Sample 2 From Data Base
Construct Score Distribution
Construct Score Distribution
Identify Quartile Cutoff Scores
Identify Quartile Cutoff Scores
Divide Into Quartiles
Divide Into Quartiles
Derive Discriminant Functions
Derive Discriminant Functions
Simulate Sequential Test
Simulate Sequential Test
Identify Quartile Placement According to Sample 2 Cutoff Scores
Identify Quartile Placement According to Sample 1 Cutoff Scores
Evaluate Accuracy of Procedure
Evaluate Accuracy of Procedure

cutoff scores for placing each examinee into one of four groups representing a quartile. In each study one sample was identified as the Developmental Sample and one sample was identified as the Validation Sample. Discriminant functions were derived employing the 256 Developmental Sample examinees in each case. The sequential testing procedure was simulated by employing the 256 Validation Sample examinees. Sample 1 was designated the Developmental Sample and Sample 2 was designated the Validation Sample for Study 1. In the second study, Samples 1 and 2 were designated the Validation and Developmental Samples respectively. Thus, each sample was alternately employed as either a validation or a developmental sample. Figure 1 depicts this procedure graphically. Table 2 presents the sample sizes and inclusive scores for each sample in each of the two studies.

The discriminant functions were derived to classify the 256 Developmental Sample examinees into four groups representing quartile placement based on the full 48 items of the test. Forty-four sets of discriminant functions were computed. Functions were derived for four variables (scores for Items 1, 2, 3, 4, 5), and in like manner, for all items through Item 48. At each stage, four sets of coefficients were computed representing the four groups to which an individual could be assigned. The coefficients generated were retained for later use by the sequential testing simulation employing the validation subjects.

Successive simulations were performed using the 45 sets of discriminant functions derived from the Developmental group's responses. The first four items were scored and converted to linear combinations using the set of coefficients for the first four items in the test. The posterior probability of membership in each group was computed and retained for later analysis. The first five test item scores were then converted to linear combinations and evaluated in a similar fashion. The process was repeated until all 48 items were included. At each step in the procedure, the probabilities of group membership were computed and retained. As each examinee's total score on the test was known, the "true group" membership was available for assessment of the accuracy of each classification. Each of the simulations concluded with a compilation of each examinee's classification history, a summary of the frequency of correct and incorrect classifications, and a summary of the average

Table 2

Sample Sizes of Developmental and Validation Groups for Studies 1 and 2

Group	Inclusive Scores	Sample Size	
		Sample 1	Sample 2
		STUDY 1	
		Developmental	Validation
1	1-21	56	71
2	22-29	68	85
3	30-35	63	42
4	36-48	69	58
Total		256	256
		STUDY 2	
		Validation	Developmental
1	1-20	49	60
2	21-26	52	67
3	27-33	67	58
4	34-48	88	71
Total		256	256

Table 3

Percent Correct Classifications and Mean Number of Items Required for Six Termination Levels in Studies 1 and 2

Term. Level	STUDY 1			STUDY 2		
	Number Classified	Percentage Correct	Mean Items Required	Number Classified	Percentage Correct	Mean Items Required
			O V E R A L L			
.70	256	62	12.00	256	66	13.45
.75	255	64	13.80	253	70	15.52
.80	255	67	16.64	247	74	17.90
.85	249	71	18.34	237	81	20.75
.90	240	74	21.69	225	85	23.77
.95	209	83	25.60	193	90	27.31
			G R O U P 1			
.70	71	73	7.82	49	86	7.29
.75	71	75	8.44	49	90	8.74
.80	71	79	9.28	49	90	9.81
.85	70	80	10.01	48	94	11.27
.90	70	77	12.57	46	91	11.93
.95	63	86	14.52	43	95	14.51
			G R O U P 2			
.70	85	54	14.06	52	48	14.98
.75	85	55	16.00	51	53	17.06
.80	85	59	19.02	47	64	19.89
.85	83	65	20.82	42	74	23.47
.90	78	72	24.82	40	80	28.05
.95	67	82	29.04	29	79	33.28

Table 3 (cont)

Percent Correct Classifications and Mean Number of Items Required for Six Termination Levels in Studies 1 and 2

	STUDY 1			STUDY 2		
Term. Level	Number Classified	Percentage Correct	Mean Items Required	Number Classified	Percentage Correct	Mean Items Required
GROUP 3						
.70	42	40	16.86	67	33	17.69
.75	41	44	19.88	65	40	20.00
.80	41	49	22.56	63	44	23.98
.85	39	56	26.00	60	57	28.96
.90	35	63	28.17	54	69	33.56
.95	29	76	33.69	40	77	38.08
GROUP 4						
.70	58	76	10.59	88	91	12.75
.75	58	79	12.86	88	91	15.10
.80	58	79	17.97	88	93	16.99
.85	57	81	19.70	87	94	18.99
.90	57	79	24.63	85	95	21.93
.95	50	84	30.26	81	98	26.65

number of items required to classify the examinees under each of six termination levels. These summaries provided the primary basis for evaluation of the discriminant function approach to sequential testing.

Results

The sequential testing procedure was evaluated by inspection of the extent to which examinees were accurately assigned to the quartile in which they would be assigned on the basis of their total score. Item economy was also an evaluation factor. Overall and quartile-group tabulations of the simulation results from both studies are presented in Table 3. The table shows that for Study 1 the number classified was less than the total sample for termination levels exceeding .70 overall, and exceeding .80, .80, .70, and .80 for groups 1, 2, 3, and 4 respectively. For Study 2 the corresponding termination levels were .70, .80, .70, .70, and .80 for the entire sample and the four respective groups. The difference between the total number classified and the total sample size (Validation Sample sizes indicated in Table 2) represents those examinees for whom all 48 items were "administered" in the simulation. Data for those unclassified examinees are not represented in the computation of the results presented.

Examination of Table 3 reveals that 62% to 83% of those examinees classified were classified correctly in Study 1. Sixty-six to 90 percent were correctly classified in Study 2. A somewhat higher accuracy rate was found in Study 2 at all termination levels. Differential accuracy levels were found for the four quartile-groups in both Study 1 and Study 2. In both studies, higher accuracy was shown in groups 1 and 4 than in groups 2 and 3. In the majority of instances, the percentage correctly classified increased as the termination level increased. The number of examinees classifed overall and in each group decreased with the increase in termination level. Eighteen percent of the examinees in Study 1 and 25 percent of those in Study 2 failed to be classified by the model at the .95 termination level. At the more moderate .80 termination level, virtually all (99% and 96% for Studies 1 and 2 respectively) examinees were classified, approximately 70% correctly. Kolmogorov-Smirnov one sample tests were performed to test whether the obtained accuracy levels were those obtainable from a population with equal accuracy across groups. The tests were repeated for each termination level in the two studies. The tests showed no evidence of departure from equal accuracy levels ($p > .10$) at any termination level in Study 1 and for termination levels of .80 and above in Study 2. Significant differences from distribution congruence were indicated for termination levels of .70 ($p < .01$) and .75 ($p < .05$).

The procedure required an average of 12 to 26 items from the 48 item test in Study 1 for the 6 termination levels, and 14 to 28 items in Study 2. Differential item ecomony was shown for the four ability levels in both studies. For nearly all termination levels the ranked mean number of items required was Group 3-Group 2-Group 4-Group 1, with Group 3 requiring the highest mean number of items for classification, more than double the mean number required for Group 1.

The high error rate suggested the analysis depicted by Table 4. The table indicates that the majority of errors resulted from assignment of an examinee to an ability level one level removed from the true level (degree of error:1). Thus, at a termination level of .75, for example, all errors for Group 3 of Study 1 (56.1% of the examinees) were for classifications to either Group 4 or Group 2; no errors occurred by assigning an examinee to Group 1. Table 5 further explicates the nature of the classification errors. A termination level of .80 was selected for further analysis of the classification errors. For each error of classification, the difference between the examinees' total test score and the minimum or maximum score required for the assigned group was computed as a measure of the extent of the error. This error score is referred to as the Boundary Score Deviation in Table 5. The Boundary Score Deviation may be described as follows: If an examinee in Group 3 (Study 1) has a total test score of 30 and was classified in Group 2, the boundary score deviation is 30-29=1, as a total score of 29 would have placed the examinee in Group 2. An assignment to Group 4 would yeild a boundary score deviation of 36-30=6 as the minimum score for Group 4 was 36.

Table 5 shows, as would be expected, that the majority of classifcation errors occurred for examinees obtaining total scores close to the boundaries of the groups in which

Table 4

Frequency and Percentage of Examinees Misclassified by 1, 2, and 3 Categories

GROUP	DEGREE OF ERROR	STUDY 1			STUDY 2		
		N	FREQ	%	N	FREQ	%
	TERMINATION LEVEL .70						
1	1	71	13	18.3	49	6	12.0
	2	71	1	1.4	49	0	-
	3	71	5	7.0	49	1	2.0
2	1	85	32	37.6	52	25	48.1
	2	85	7	8.2	52	2	3.8
3	1	42	24	57.1	67	40	59.7
	2	42	1	2.4	67	5	7.5
4	1	58	10	17.2	88	8	9.1
	2	58	3	5.2	88	0	-
	3	58	1	1.7	88	0	-
	TERMINATION LEVEL .75						
1	1	71	13	18.3	49	5	10.2
	2	71	1	1.4	49	0	-
	3	71	4	5.6	49	0	-
2	1	85	32	37.6	51	24	47.1
	2	85	6	7.1	51	0	-
3	1	41	23	56.1	65	35	53.8
	2	41	0	-	65	4	6.2
4	1	58	10	17.2	88	8	9.1
	2	58	1	1.7	88	0	-
	3	58	1	1.7	88	0	-

Table 4 (cont)

Frequency and Percentage of Examinees
Misclassified by 1, 2, and 3 Categories

GROUP	DEGREE OF ERROR	STUDY 1			STUDY 2		
		N	FREQ	%	N	FREQ	%
TERMINATION LEVEL .80							
1	1	71	12	16.9	49	5	10.2
	2	71	0	-	49	0	-
	3	71	3	4.2	49	0	-
2	1	85	32	37.6	47	17	36.2
	2	85	3	3.5	47	0	-
3	1	41	21	51.2	63	31	49.2
	2	41	0	-	63	4	6.3
4	1	58	11	19.0	88	6	6.8
	2	58	0	-	88	0	-
	3	58	1	1.7	88	0	-
TERMINATION LEVEL .85							
1	1	70	12	17.1	48	3	6.3
	2	70	0	-	48	0	-
	3	70	2	2.9	48	0	-
2	1	83	26	31.3	42	11	26.2
	2	83	3	3.6	42	0	-
3	1	39	17	43.6	60	24	40.0
	2	39	0	-	60	2	3.3
4	1	57	10	17.5	87	5	5.7
	2	57	0	-	87	0	-
	3	57	1	1.8	87	0	-

Table 4 (cont)

Frequency and Percentage of Examinees Misclassified by 1, 2, and 3 Categories

GROUP	DEGREE OF ERROR	STUDY 1			STUDY 2		
		N	FREQ	%	N	FREQ	%
TERMINATION LEVEL .90							
1	1	70	15	21.4	46	4	8.7
	2	70	0	-	46	0	-
	3	70	1	1.4	46	0	-
2	1	78	19	24.4	40	8	20.0
	2	78	3	3.8	40	0	-
3	1	35	13	37.1	54	16	29.6
	2	35	0	-	54	1	1.9
4	1	57	12	21.1	85	4	4.7
	2	57	0	-	85	0	-
	3	57	0	-	85	0	-
TERMINATION LEVEL .95							
1	1	63	9	14.3	43	2	4.7
	2	63	0	-	43	0	-
	3	63	0	-	43	0	-
2	1	67	11	16.[illegible]	29	6	20.7
	2	67	1	1.5	29	0	-
3	1	29	7	24.1	40	8	20.0
	2	29	0	-	40	1	2.5
4	1	50	8	16.0	81	2	2.5
	2	50	0	-	81	0	-
	3	50	0	-	81	0	-

Table 5
Frequency and Percentage of Examinees Misclassified in Each Group by Boundary Score Deviation for Termination Level = .80

Study	Boundary Score Deviation	GROUP 1 F	GROUP 1 %	GROUP 2 F	GROUP 2 %	GROUP 3 F	GROUP 3 %	GROUP 4 F	GROUP 4 %
ONE DEGREE ERROR									
1	1	8	11	6	7	5	12	5	9
	2	3	4	10	12	6	15	4	7
	3	0	0	5	6	1	2	1	2
	4	1	1	5	6	1	2	1	2
	5	0	0	4	5	4	10	0	0
	6	0	0	1	1	4	10	0	0
	7	0	0	1	1	0	0	0	0
	Total	12	17	32	38	21	51	11	19
2	1	4	8	1	2	13	21	3	3
	2	0	0	8	17	5	8	2	2
	3	0	0	2	4	6	10	0	0
	4	1	2	2	4	2	3	0	0
	5	0	0	2	4	2	3	0	0
	6	0	0	2	4	0	0	0	0
	7	0	0	0	0	1	2	0	0
	Total	5	10	17	36	31	49	6	7
TWO DEGREE ERROR									
1	7	0	0	2	2	0	0	0	0
	8	0	0	0	0	0	0	0	0
	9	0	0	0	0	0	0	0	0
	10	0	0	0	0	0	0	0	0
	11	0	0	0	0	0	0	0	0
	12	0	0	1	1	0	0	0	0
	Total	0	0	1	1	0	0	0	0
2	7	0	0	0	0	0	0	0	0
	8	0	0	0	0	1	2	0	0
	9	0	0	0	0	1	2	0	0
	10	0	0	0	0	1	2	0	0
	11	0	0	0	0	1	2	0	0
	12	0	0	0	0	0	0	0	0
	Total	0	0	0	0	4	6	0	0

they were classified by the procedure. Boundary score deviations of 3 points or less account for over 90% of the errors in Groups 1 and 4, and over 50% of the errors in Groups 2 and 3. The errors are again shown to be of greater magnitude for examinees in Groups 2 and 3 in both studies.

Discussion

Evaluation of the sequential testing procedure was performed using two criteria: classification accuracy and item economy. Classification accuracy is indicated by the percentages given in Table 3 and 4. The most striking result of the analysis is the apparent differential classification accuracy from group to group. This finding is not at all clear in view of the results of the Kolmogorov-Smirnov tests which in general did not reveal significant variance from equal group to group accuracy. The nature of the K-S tests leads this researcher to suspend judgment on the differential accuracy question pending review of additional data. It may be reported, however, that the findings of two other studies by the author, lacking K-S tests, have indicated differential accuracy of the same direction and magnitude of that indicated in the present studies.

As Table 4 indicates, major classification errors occur for about 5% of the examinees, primarily those in Groups 2 and 3. The review of individual examinee classifications reported in Table 5 indicates that the incorrect classifications of a single category occur for those examinees whose total scores lie near the quartile boundaries (e.g., an examinee with a total score of 21 may be classified incorrectly in Group 2, the lowest score for which was 22 in Study 1). Thus, the classification errors are generally not major ones.

An important question is, "How accurate does the test have to be to adequately meet an instructional need?" The answer may lie in an analysis of the variety of instructional alternatives at any decision point. If three alternatives are available, then a testing procedure which differentiates between those students for whom alternative A is appropriate and those for whom alternatives B or C are more appropriate should be sufficient. It appears that the classificaton errors are not of sufficient magnitude as to preclude use of the procedure for the purpose described.

It may be concluded that the accuracy of the sequential testing procedure is probably adequate for many instructional classification decisions. Those examinees who were incorrectly classified by one quartile group, were not badly served by the procedure, as their total scores very nearly would have placed them into the assigned group. Were instruction to follow at a higher or lower level (pace, content, difficulty) any significant imprecision would doubtless be corrected through subsequent monitoring and analysis of the student's progress.

A review of Table 3 indicates that a substantial reduction in the number of items required for classification is possible in comparison to the number of items in the full test. For examinees in Groups 1 and 4, classification may be made with an average of 1/3 the items in the full test. The middle groups require approximately 1/2 the items. Thus, is is concluded that item economy is significant. The time saving could amount to 10-12 minutes per administration. Stated differently, twice the number of examinees could be administered a long series of tests in a given time period.

Certain examinee's responses could not result in classification for termination levels higher than .80. A procedure is required to provide a classification for these otherwise non-classified examinees. One procedure is to administer all items to those few non-classified examinees, if necessary, and to transform the usual total score into the quartile assignment. Another procedure would be to retain a number correct score and continue testing until sequential classification had succeeded or until the total number correct at Item N plus the sum of all future items (assumed correct) fell within the same range of the classification current for the total number correct at Item N (e.g., if 44 items had been administered from a 48 item test, and 31 were correct, testing could terminate if sequential classificaton had not occurred, as 31 and [48-44]+31 are both within the range of Group 3 [see Table 2, Study 1]). The effect of such procedures has not been evaluated. Such modifications to the method presented here are certainly possible and would increase accuracy somewhat for the few examinees affected.

Cost and time have long been important factors in the design and administration of military training. The cost of computer technology has been dropping rapidly and is currently available to many sectors of the educational enterprize, but the relative cost and effectiveness of computer-based instruction and testing is still of concern. The developmental cost of computer-based tests and the time (both human and computer) to administer them at a terminal is in many cases prohibitive. The use of adaptive tests, whether branching or sequential, will permit a more comprehensive array of tests to be administered by computer within a short time period. When existing tests are prefereable, time and cost a consideration, and computer administration desirable, the sequential procedure would be preferred over the branching procedure. This research is not conclusive as research which includes actual computer terminal administration and scoring has yet to be performed using the technique, but the results are sufficiently encouraging that research should continue on the technique, especially for developing a means for improving classification accuracy. Item economy and improved accuracy would permit one the time to measure many of the learner characteristics necessary for a prescriptive learning model.

References

Cleary, T. A., Linn, R. L., & Rock, D. A. Reproduction of total test score through the use of sequential programmed tests. Journal of Educational Measurement, 1968, 5, 183-187.

Hieronymus, A. N., & Lindquist, E. F. Iowa tests of basic skills, Forms 5 and 6: Manual for administrators, supervisors and counselors. Boston: Houghton Mifflin Company, 1974.

Tatsuoka, M. M. Multivariate analysis: Techniques for educational and psychological research. New York: John Wiley and Sons, Inc., 1971.

Wald, A. Sequential analysis. New York: John Wiley and Sons, Inc., 1947.

Weiss, D. J. Strategies of adaptive ability measurement (Research Report 74-5) Minneapolis, MN: Department of Psychology, Psychometric Methods Program, December, 1974.

Wood, R. Response contingent testing. Review of Educational Research, 1973, 43(4), 529-544

Computerized Adaptive Ability Measurement[1]

David J. Weiss
University of Minnesota

Adaptive Testing and Error Reduction

The general objective of our research program on adaptive testing is to identify several sources of potential error in test scores, and to study adaptive testing as a means for reducing these errors of measurement.

The first general source of error that we have been concerned with is the error that results from the mis-match of item difficulties in an ability test with the individual's ability. Obviously, the testee's ability is not known at the start of testing. But the different strategies of adaptive testing that have been proposed can be viewed as different ways of matching item difficulties with testee ability and sequentially estimating the testee's ability. Consequently, one of our major emphases is to determine the best, or at least better, ways of adapting item difficulties to individual abilities. Much of what I have to say today will be concerned with these various strategies of adaptive testing.

We are approaching this in two complementary ways. First, we have been doing live computerized testing. Since late 1972 we have tested more than 5,000 subjects on a variety of strategies of adaptive testing. But live testing cannot provide answers to all the questions concerning which strategies are best under which conditions, because there are too many questions to be answered. Therefore, we are using computer simulation to supplement and extend the results that we obtain from live testing.

The second main emphasis of our research is a concern with the psychological effects of adaptive testing. Here we are concerned with identifying the psychological aspects of testing and the test environment which can introduce error into test scores. These variables include guessing, test anxiety, boredom, frustration, lack of motivation, and racial or ethnic group effects.

---

[1] Early development work on this research was supported during 1969 and 1970 by grants from the General Research Fund of the Graduate School, University of Minnesota. Research reported in this paper was supported since early 1972 by Personnel and Training Research Programs, Office of Naval Research, Contract No. N00014-67-A-0113-0029, NR 150-343.

Portions of this paper were written by the project staff: Nancy Betz, Jim McBride, Brad Sympson and Dave Vale.

Special thanks are due to our project programmer, John Dewitt, without whom this research would have been nearly impossible.

Guessing can obviously artificially increase test scores; frustration, anxiety, motivation and other factors can result in test scores lower than true ability. All of these, therefore, are sources of error in test scores which are due to the psychological effects of testing.

We are also concerned with the psychological effects that will result from the man-machine interface. This, from our experience, is going to be an important problem in computerized adaptive testing. There are different kinds of computer systems on which we can implement adaptive testing and each of those computer systems has its positive and negative effects on testee behavior. There are different kinds of terminal devices for adaptive testing and each kind of terminal device displays in different ways and at different speeds. All of these variations in the man-machine interface are going to be new problems for us to consider in the years to come. Past research has demonstrated that answer sheets in paper and pencil testing sometimes had an effect on test scores. Similarly, research in adaptive testing will need to study different kinds of CRT's, different kinds of computer systems and different display speeds as part of the psychological effects of computerized testing. In the second half of today's presentation I will present some data relevant to the psychological effects of adaptive testing.

A third source of error that we are concerned with is error that results from not extracting enough information from a testee's response to a test item. To date most psychometric research has been concerned with binary or 0-1 scoring. But we can extract more information from a test response if we assign different scores to different incorrect response alternatives. Test responses can be even more informative if we use continuous responding, or probabilistic responding.

The fourth source of error that we are studying is the error that results from deviations from unidimensionality. Latent trait theory, as it is usually used in testing, is based on the assumption of unidimensionality, although there are multidimensional latent trait models being developed. But dimensionality that is defined on a group, such as the unidimensionality of latent trait theory, does not necessarily hold true for an individual. That is, dimensionality defined by factor analysis or other methods, when applied to an individual, assumes that the individual is the typical or average member of the group on which the dimensionality was defined. Thus, in the testing situation, when a set of "unidimensional" items is administered to an individual, the result may be a set of responses that are not unidimensionally determined.

Consequently, our research is concerned with individual-item pool interactions--the interaction of one individual with a set of "unidimensional" items. We are studying item response protocols of this nature to determine if meaningful deviations from unidimensionality do occur for specific individuals. If they do, we will then develop interactive adaptive testing models that will take account of intra-individual multidimensionality. I'll have more to say about the dimensionality problem later.

A fifth kind of error that we plan to study in the future is the error that results from an over-simplistic conceptualization of ability. In the past fifty years, we have largely let the nature of our ability tests be determined by the restrictions imposed by the paper-and-pencil testing medium. Thus, many of our abilities are "static" abilities, such as verbal ability measured by the multiple-choice vocabulary test. But interactive computer systems permit us to break out of these shackles and measure abilities that are not measureable in paper-and-pencil formats. We should now be able to measure such abilities as reasoning, by following an individual's chain of decisions given a structured set of problem stimuli. Or, we will be able to measure memory abilities within a dynamic framework, or perceptual abilities, including perception of movement, using computer-controlled stimuli. The possibilities are endless, and the net result should be new kinds of ability measures which will likely be more meaningful and accurate for occupational prediction.

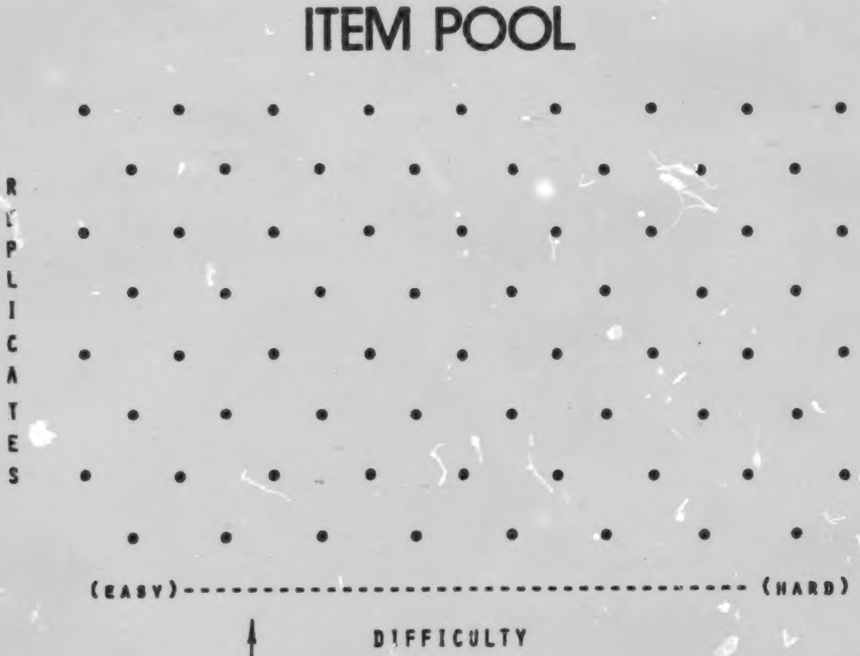
Strategies of Adaptive Testing

The bulk of our research during the last several years has been concerned with the first type of error. That is, we have been studying various methods for selecting items from a pre-calibrated pool, to match each individual's ability level as it is estimated during the process of testing. The basic premise of adaptive testing is this: an individual's ability level will be most accurately estimated when the items administered are as close to his/her ability level as possible. But, since ability is not known before testing--since that is the purpose of administering the test--we must choose items for each individual while testing is in progress. Thus, computerized adaptive testing uses an interactive computer system to administer tests; each item is chosen based on the testee's responses to previous items. Items are typically administered on a cathode-ray-terminal (CRT), and the testee responds on the CRT keyboard.

I will describe some strategies that have been used for selecting items in the framework of their evolution from the simple conventional test to complex adaptive or tailored testing models. To clarify the distinctions between some of the models we will follow the progress of a hypothetical low ability subject through a test administered under each strategy and note how his items are selected. We will further examine differences between strategies.

Figure 1 shows the item pool that will be used to describe the way the various testing strategies function. On the horizontal dimension we have 17 columns, each containing four items, ranging from very easy items at the left to very difficult items at the right. The vertical dimension represents replications of items at each difficulty level; all items in a column are equally difficult.

Figure 1



ITEM POOL

91

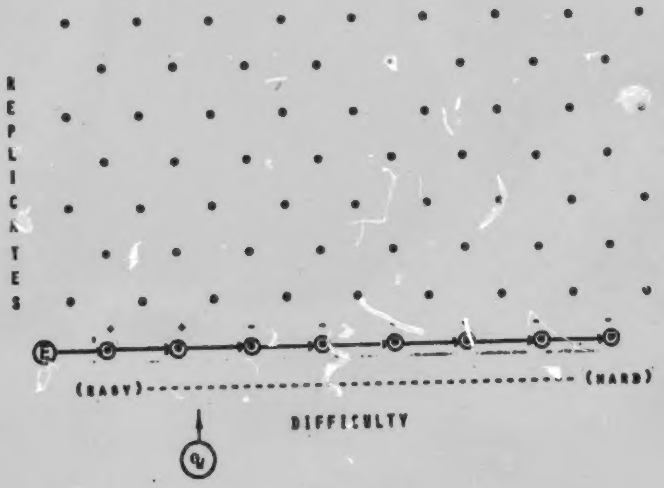
I will illustrate the various item selection strategies using eight items from this pool of 68. While an eight item test is convenient for illustration, eight items are too few for measurement of reasonable accuracy. Therefore, for evaluation of the strategies a 24-item test was used. Items for the 24-item test were chosen in a manner analogous to the way items were chosen for the illustrated eight-item test.

The results that I will present are from computer simulations. In order to make possible the analyses done for this presentation, some simplifying assumptions were made. First, it was assumed that a large pool of equally good items (i.e., items with equivalent discriminating power) was available to chose from. Second, it was assumed that these were free-response items and, hence, guessing was not possible. Third, it was assumed that all tests were scored by a common technique, in this case, a Bayesian scoring procedure. Finally, to make comparisons between some strategies meaningful, it was assumed that a prior estimate of ability, correlating 0.5 with ability, was available.

One way to compose a test is to select a fixed set of items having a wide range of difficulties. Figure 2 shows such a rectangular conventional test. In this case, eight items equally spaced on the difficulty continuum were chosen from alternate columns ranging from the next to easiest to the next to most difficult columns. Our low ability subject, produced the response record shown, with those items he answered correctly marked by a "+" and those he answered incorrectly indicated by a "-". The items in this test could have been administered in any order, but for clarity of presentation, we started at the left and worked toward the right.

Figure 2

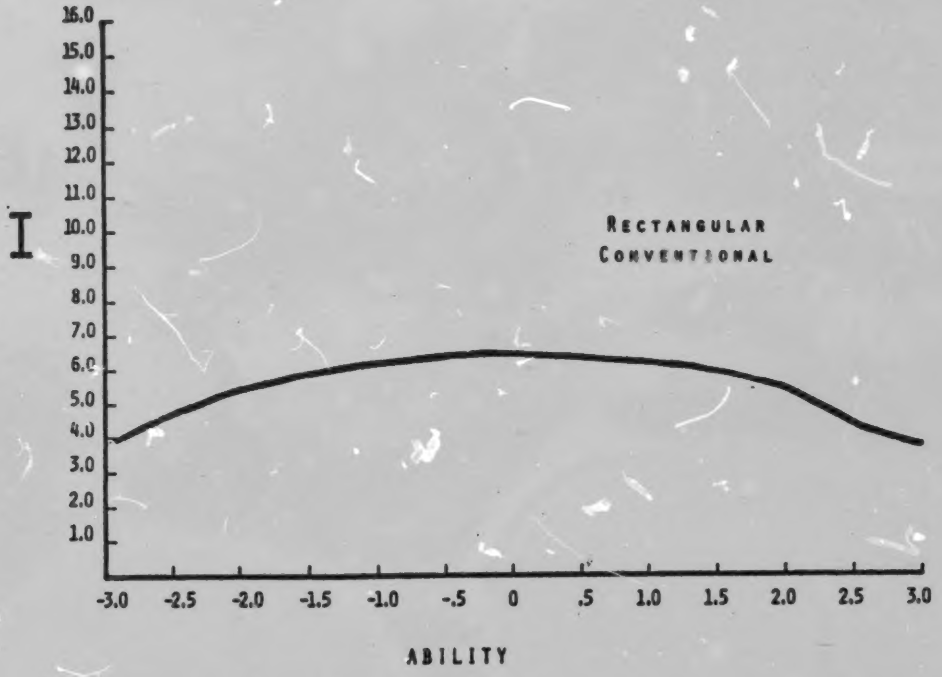
RECTANGULAR CONVENTIONAL



The first item encountered was beneath the testee's ability level (θ_d) and knowing the answer, he responded correctly. The second item was a bit more difficult but he still answered it correctly. The third item, being a bit above his ability was too difficult and he answered it incorrectly. Similarly, the fourth through eighth items were even more difficult and he answered all of them incorrectly.

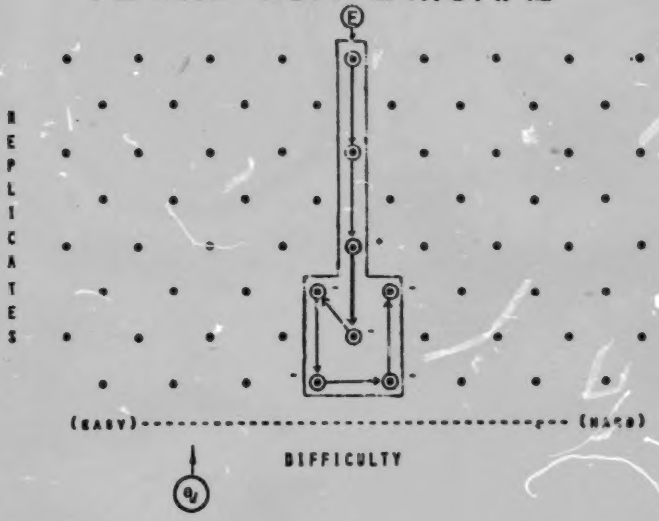
Figure 3 shows an information curve produced by the rectangular conventional test. Information can be thought of as related to the precision of measurement produced by a test at a given level of ability, or as how well a test can discriminate between two contiguous ability levels. A good test produces an information function that is high (i.e., provides precise measurement) and is flat (i.e., provides this high level of precision for all testees at all ability levels. Although not apparent from Figure 3, it will become obvious from comparisons with later results that the rectangular conventional test produces an information function that is fairly flat but somewhat low. It can be seen, however, that even this information function tapers off at the extremes indicating poorer measurement for testees where ability level is distant from the mean.

Figure 3



Instead of choosing items with a wide range of difficulty, we could instead choose items peaked at the center of the ability range and administer them to all testees. Figure 4 shows such a peaked conventional test. We chose the four items from the median difficulty column and two from each of the adjacent columns. Again, these items could have been administered in any order but we began at the top for clarity.

Figure 4

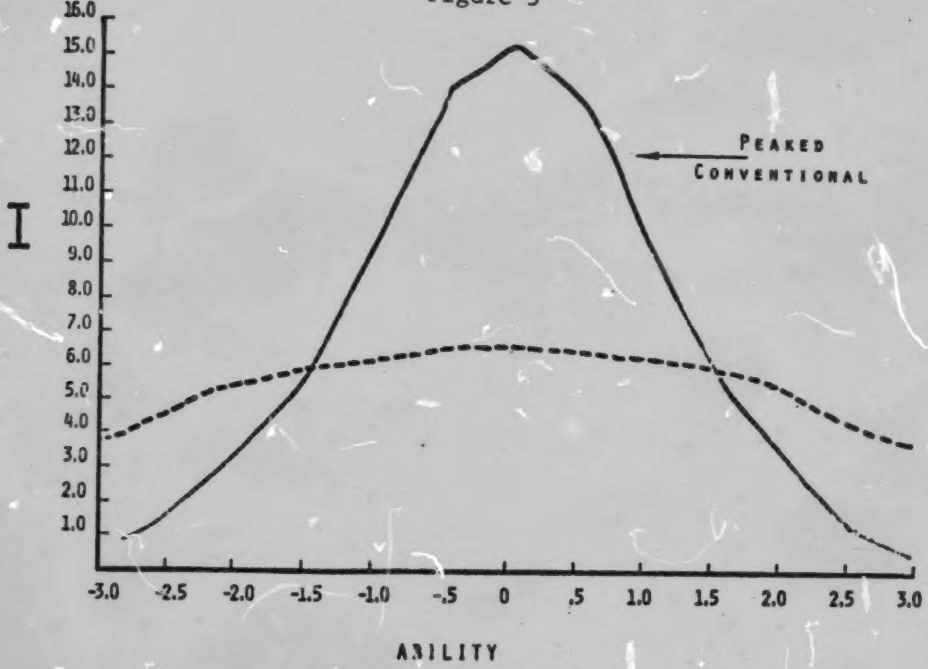


PEAKED CONVENTIONAL

These items were intended for average ability testees and were all too difficult for our low ability testee. He missed the first item, the second item, and most of the rest of the items.

The information curve for the peaked conventional test (Figure 5) shows graphically what our testee felt as he took the test; the peaked conventional test provides good measurement for some testees but very poor measurement for others. As Figure 5 shows, the peaked conventional test produces precise measurement for individuals with abilities in the middle range but little information for extreme ability subjects. The peaked conventional test provides more information about ability than does the rectangular conventional test within the range of ±1.5 standard deviations of ability but less outside of this range.

Figure 5

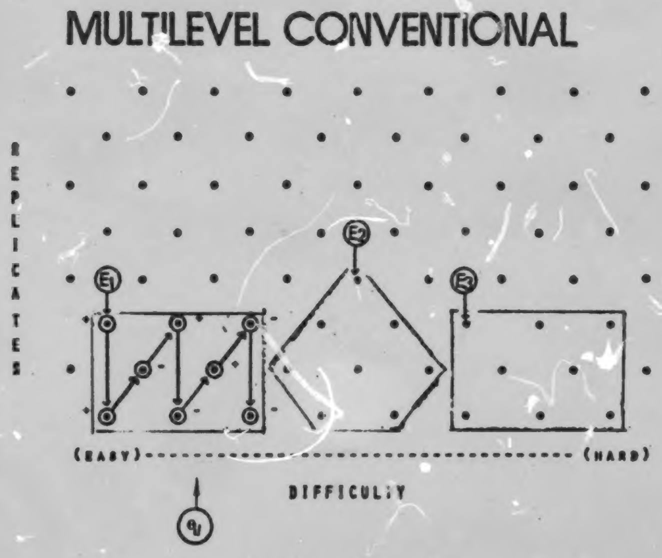


It seems that with a fixed set of items (i.e., a conventional test) we can please some of the people all of the time or all of the people some of the time but can't please all of the people all of the time. If, however, we could figure out a way to move a peaked ability test to the ability of each person being tested, we could please all of the people all of the time and provide a high level of information at all ability levels. If a testee's ability were known a priori, we would construct a test made up of those items with difficulties closest to his ability (i.e., items which he would be expected to answer correctly 50% of the time). But, if we knew his ability beforehand, we would have no reason to administer the test at all.

In practice we have, at best, a fallible prior estimate of the testee's ability and may want to administer items more or less rectangularly distributed in a narrow range around his estimated ability. Some achievement tests use a prior ability estimate, such as grade in school, to determine which section of a test a testee should take.

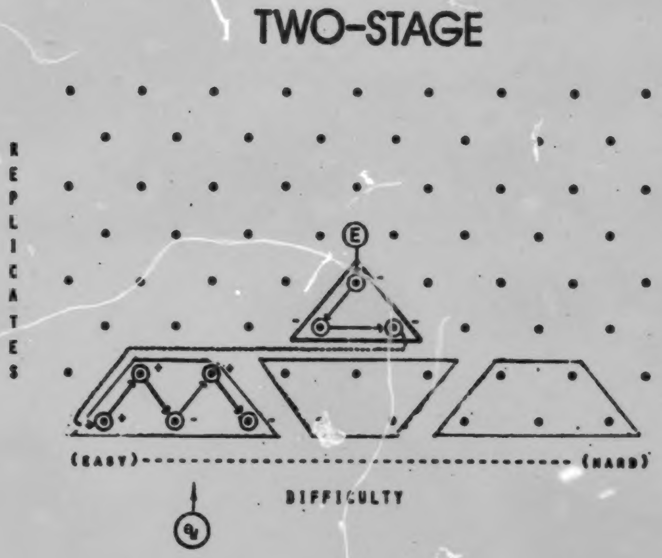
Figure 6 illustrates such a test. Knowing that a testee ranked at the 27th percentile in his grade school graduating class, if this were a high school freshman achievement test, we might use this prior information to start him at the easiest entry point (E_1). Or, if we had a testee with straight A's in grade school, we might start him at the high entry point (E_3). Given a prior ability estimate, therefore, it is possible to adapt the test to the individual within the framework of a conventional test. But if prior information is not available, we have to use a test that tailors item difficulty in its absence. One possible strategy for doing this is the two-stage testing strategy which is like the previous test but generates its own prior ability estimate.

Figure 6



In a two-stage test, a testee is first administered a short routing test and, on the basis of his score on that test, is branched to a measurement test of more appropriate difficulty. Figure 7 shows a two-stage test. A testee takes a three-item routing test and one of three five-item measurement tests. Our low ability testee answered all three of the routing test items incorrectly as they were too difficult for him. Since this suggested that his ability was low, he was branched to the easiest measurement test where he answered three out of the five items correctly.

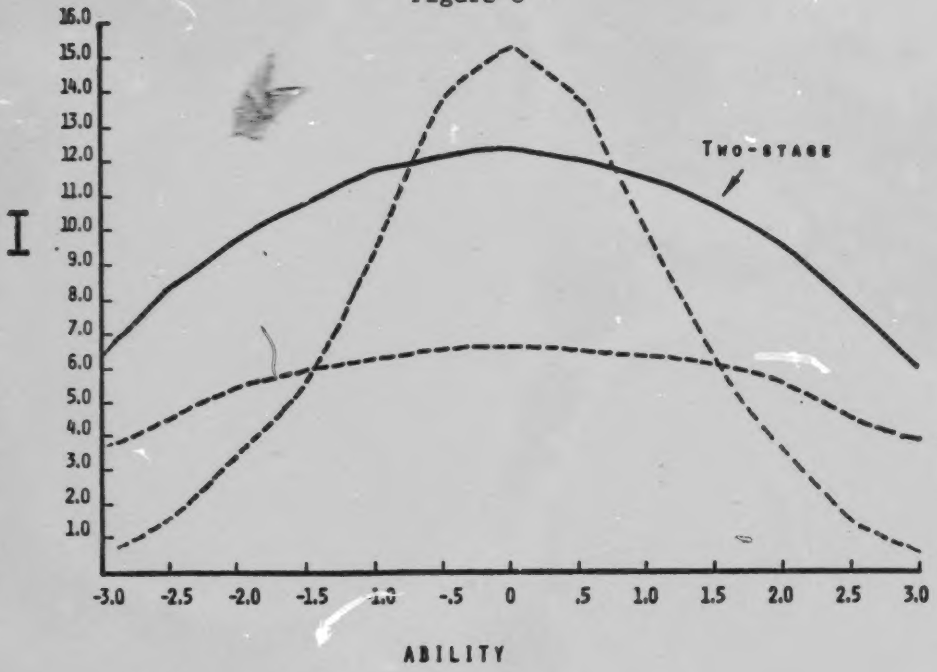
Figure 7


TWO-STAGE

As Figure 8 shows, this two-stage test yields an information curve that is at all points higher than the rectangular conventional test and higher than the information curve of the peaked conventional test except in the center. So this two-stage test provides more precise measurement than the rectangular conventional test at all ability levels and more precise measurement than the peaked conventional test at most ability levels.

One problem with the two-stage testing strategy is that if a testee's ability is between the difficulties of two adjacent measurement tests, there is no measurement test of appropriate difficulty. A solution to this problem is available in the form of

Figure 8

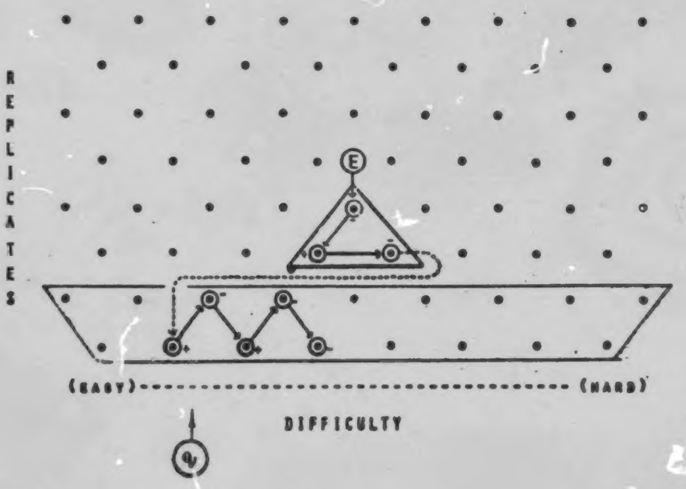


the continuous second stage two-stage test (Figure 9), a variant of the previous two-stage test. As in the standard two-stage test, the testee is first administered the routing test. Then, on the basis of the score on that test, he is branched to a measurement test. But instead of using one of a series of pre-structured measurement tests, a measurement test is individually composed for that individual using items closest to his ability estimate, plus items on either side. Given our restricted circumstances, the information curve of the continuous two-stage test would be very similar to that of the standard two-stage test and will not be shown here.

Another problem inherent in the two-stage procedure is that of misrouting. The measurement test decision is based on a short and fallible routing test and thus may be incorrect. There are two solutions to the misrouting problem: One is to route more; the other is to route less (i.e., not at all). An example of the latter strategy is the flexilevel test (Figure 10). For this test the potential item set is the same as the potential measurement test item set of the continuous two-stage test. But, rather than taking a routing test, each testee starts with the median difficulty item of the

Figure 9

CONTINUOUS TWO-STAGE



item set and following each correct response is branched to the next more difficult unadministered item. Following an incorrect response, he is branched to the next less difficult unadministered item.

In the case illustrated, the testee missed the first three items and was branched appropriately downward until he reached the third item below the median, an item slightly above his ability level. Knowing the answer, he answered it correctly and was branched to the first item above the median, which he answered incorrectly. He was branched to the fourth item below the median item and continued oscillating between easy and difficult items until he had answered eight items.

The information curve for the flexilevel test is shown in Figure 11. Although the flexilevel test solves the problem of misrouting, the

Figure 10

FLEXILEVEL

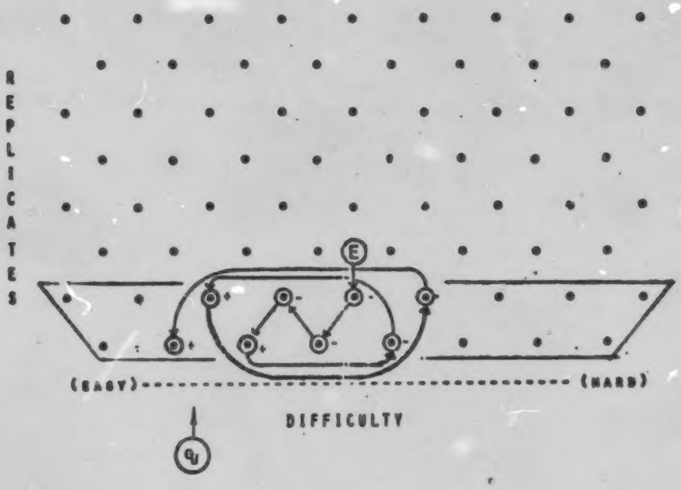
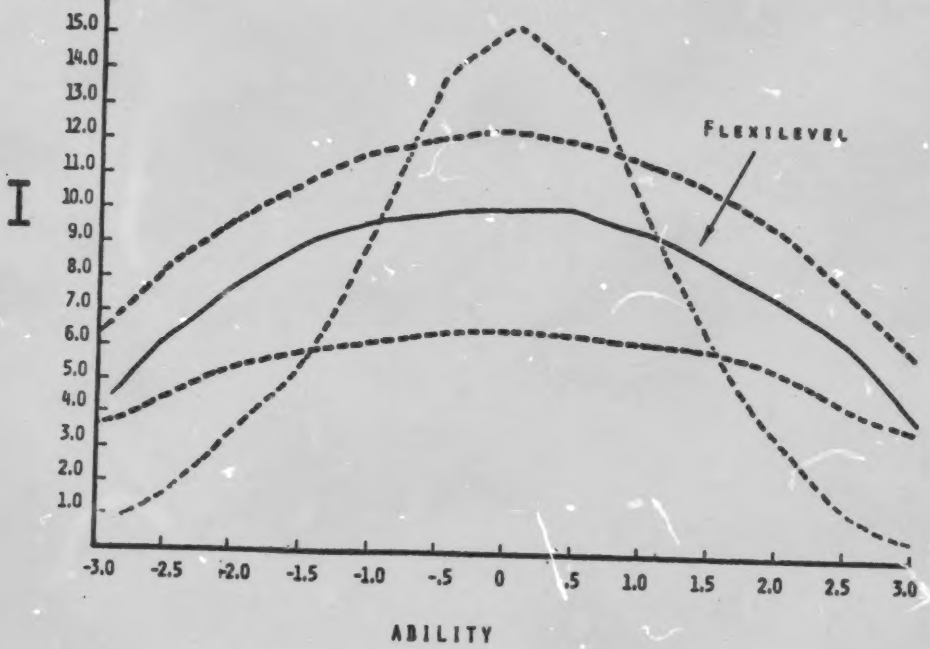

REPLICATES
E
(EASY)
(HARD)
DIFFICULTY

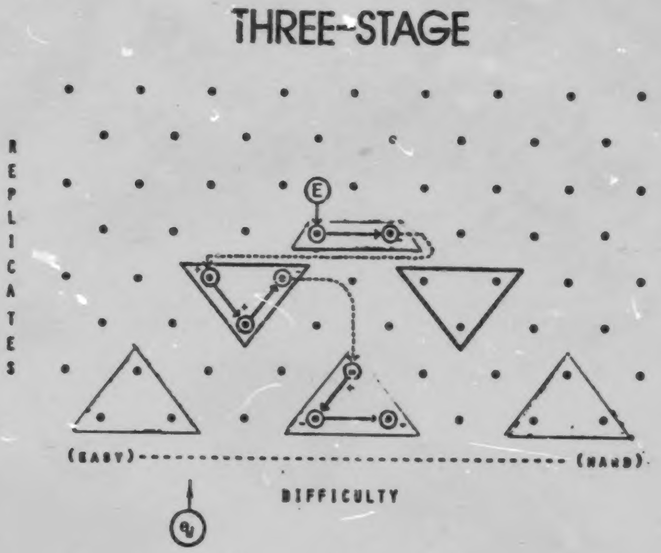
Figure 11



information it provides is always less than that provided by the two-stage test.

Figure 12 shows an example of the other solution to the problem of misrouting, the three-stage test (sometimes referred to as the double-routing two-stage test). In this strategy, an individual takes one routing test which routes him to a second routing test which routes him to a measurement test. Errors resulting from the first routing can be ameliorated by the second routing.

Figure 12


THREE-STAGE

Carrying the idea of multiple routing to its logical extreme, and using one item per stage, results, in this case, in the eight-stage test or, in the general case, the pyramidal test. In this strategy (Figure 13), a testee starts with a median difficulty item and is branched after each item. A less difficult item is administered following an incorrect response, and a more difficult item is administered following a correct response.

The information curve for this test (Figure 14) shows it to provide more information than any of the strategies discussed thus far, except in the middle ability range where it is slightly surpassed by the peaked conventional test. It should be noted, however, that the information

Figure 13

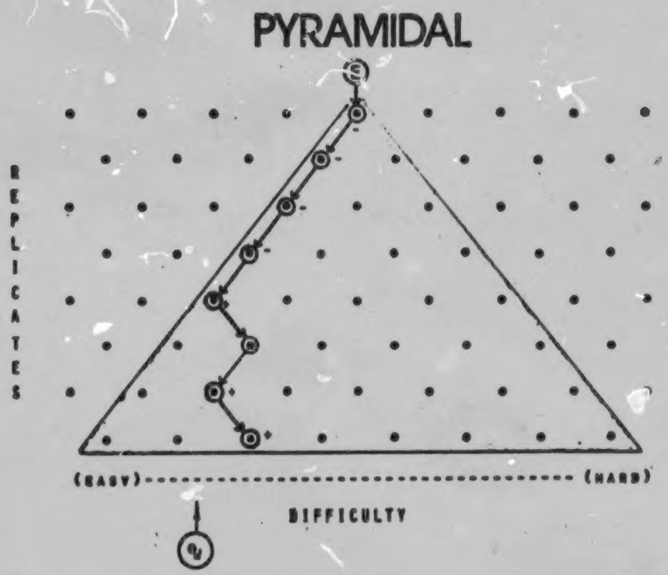
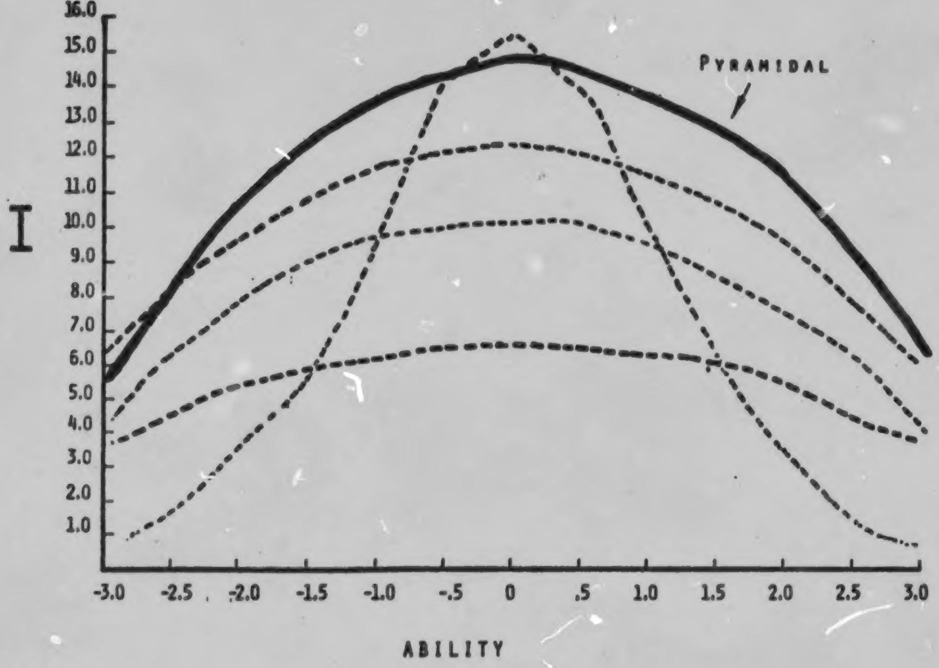

PYRAMIDAL
REPLICATES
(EASY)
(HARD)
DIFFICULTY

Figure 14

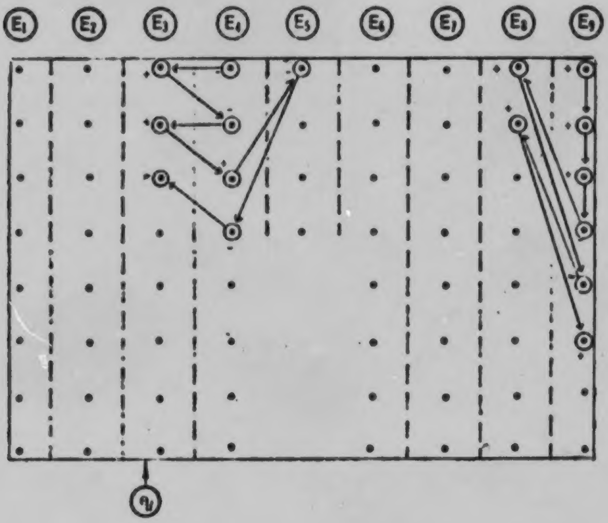


curve is far from flat. Less than half of the amount of information provided at the middle range of ability is provided at the extremes of this information curve, three standard deviations from the mean.

The previously discussed adaptive tests have been developed for the situation in which prior ability information was not available and are not capable of using it when it is available. Now that we have reached the top of the pyramid, so to speak, we can make use of prior information by extending the pyramidal structure to allow entry at several points. A direct extension is unable to handle branching for some extreme ability testees, however, so a modified extension of the pyramidal structure is used by the stratified-adaptive (stradaptive) testing strategy shown in Figure 15. Two changes beyond a direct extension are observed: 1) items are grouped into strata consisting of items of possibly slightly different difficulty; and 2) branching is between strata with the item selected being the first unadministered item in a stratum.

Figure 15

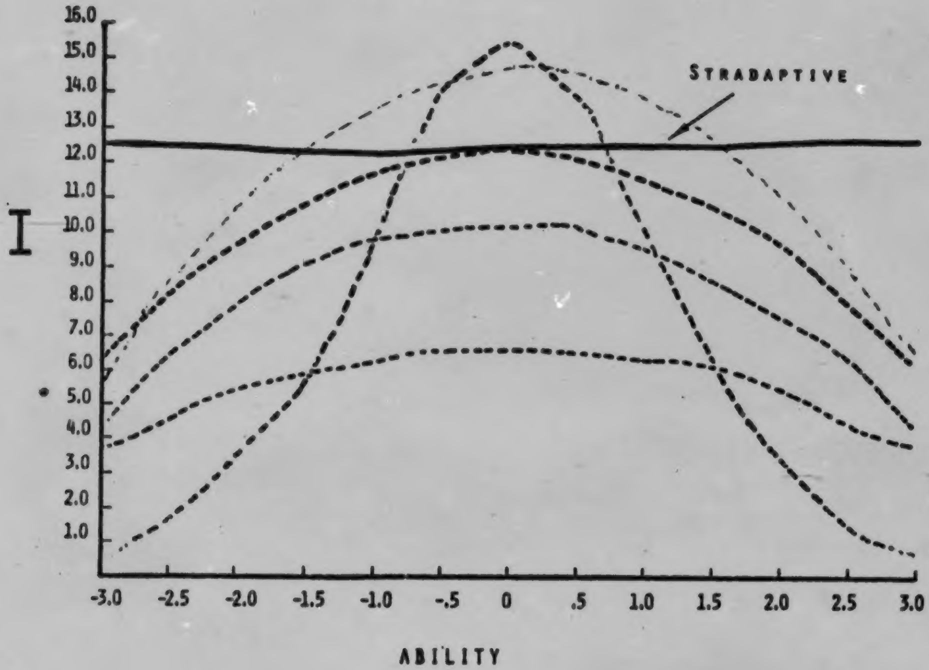
STRADAPTIVE



The testee started at the fourth entry point. He missed the first item in stratum four, was branched to the first item in stratum three, got this item correct, and alternated between these two strata until his fifth item. He answered the fifth item, which was in the fourth stratum, correctly and was branched to the first item in the fifth stratum. He incorrectly answered this and the next item and finished with his eighth item in the third stratum.

Branching to the first item in a stratum is of little value in a situation where all items are equally discriminating, but is useful when using a real item pool because all items will not be equally discriminating. This feature allows the most discriminating items to be put where they have the highest probability of being administered; as the first items to be administered in each stratum. The information curve for the stradaptive test (Figure 16) is almost flat indicating that it provides very equiprecise measurement. Its level is surpassed by several other strategies in the center, however.

Figure 16



The previous adaptive strategies are all among the fixed branching strategies. The branching has been a function solely of the testee's performance at the immediately preceeding stage. The variable branching procedures calculate an ability estimate after each item and select as the next item the item best suited for an individual of that ability.

An example of the variable branching procedures is the Bayesian strategy, which is illustrated in Figure 17. On the basis of a prior ability estimate, which may be simply the mean ability of the population of testees, a first item is selected. On the basis of the response to that item and a prior ability distribution, which may consist simply of population parameters, a score is calculated and on the basis of that score, another item is selected. This procedure is repeated, each time selecting the one item in the pool which is closest in difficulty to the last ability estimate.

Figure 17

BAYESIAN

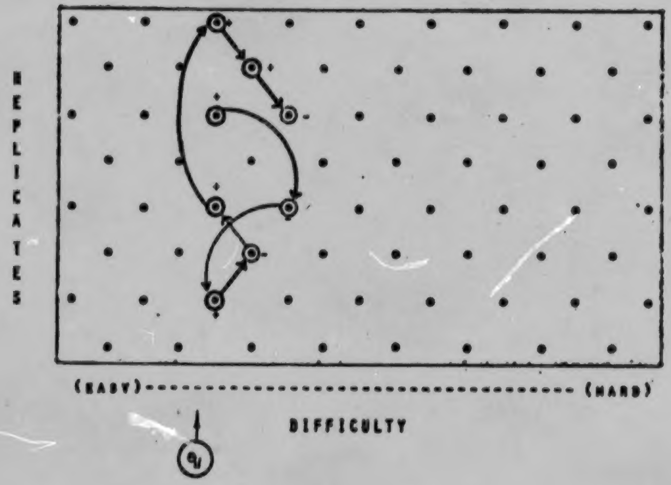


Figure 18 shows a report derived from live testing with a Bayesian adaptive test. In this figure, an "X" indicates a correct response to an item, while an "O" indicates an incorrect response; the "E" indicates the entry ability estimate, based on a rough prior estimate of the testee's ability level. The dotted lines on either side of these symbols indicate the standard deviation of the ability estimate, a value analagous to the standard error of measurement for that ability estimate. Note how the ability estimate itself (i.e., the E,X, or O) changes after each item response. Note also that the range of change in the ability estimate decreases as testing proceeds. This illustrates the convergence nature of the Bayesian process. Similarly, the error of the ability estimate decreases after each item response, with the amount of decrease reducing at each stage of the testing procedure.

Figure 18

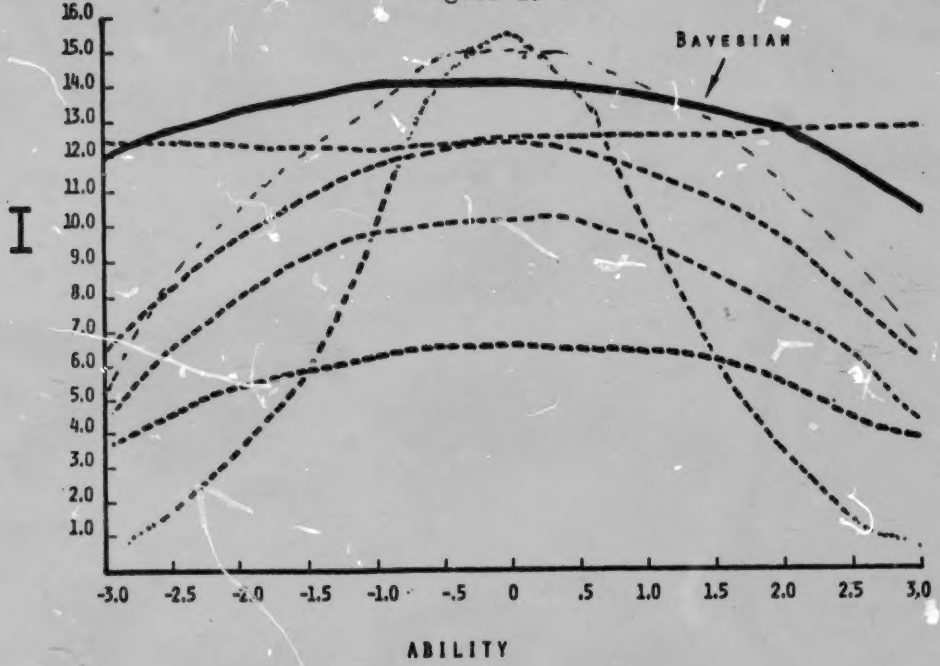
REPORT ON BAYESIAN TEST

```
               X-CORRECT   0-INCORRECT      ?-NO RESPONSE .
           ERROR BAND PLOTTED IS + AND - STANDARD DEVIATION
        LOW               ABILITY LEVEL           HIGH   POSTERIOR
          -2.5 -2.0 -1.5 -1.0 -0.5 0.0  0.5  1.0  1.5  2.0 ABILITY
STAGE I----I----I----I----I----I----I----I----I----I----I   EST    SD
  0      ..................E..................             -.85  1.87
  1                     .............X.............         .19  1.30
  2               .........0.........                       -.85   .94
  3                      .......X.......                    -.35   .76
  4                 . ......0......                         -.73   .66
  5                ......0......                           -1.07   .65
  6            .....0.....                                 -1.59   .50
  7              ....X....                                 -1.42   .44
  8                 ...X...                                -1.29   .40
  9                  ...X...                               -1.20   .39
 10                 ...0...                                -1.30   .36
 11                ...0...                                 -1.40   .35
 12                ...X...                                 -1.31   .35
 13                 ...X...                                -1.24   .33
 14                  ...X...                               -1.17   .33
 15                 ...0...                                -1.23   .32
 16                 ...0...                                -1.28   .32
 17                ...0...                                 -1.35   .30
                     17 ITEMS WERE ADMINISTERED
```

The information curve from the Bayesian testing procedure is shown in Figure 19. It is slightly higher than the stradaptive test's information curve and nearly as flat, although it drops more in the tails. The peaked conventional test and the pyramidal test still provide more information in the center of the ability distribution.

If the evaluation of adaptive testing strategies were as simple as this presentation, however, our research would be unnecessary. This evaluation was very limited in a number of ways, which seriously restrict the generalizability of these findings.

Figure 19

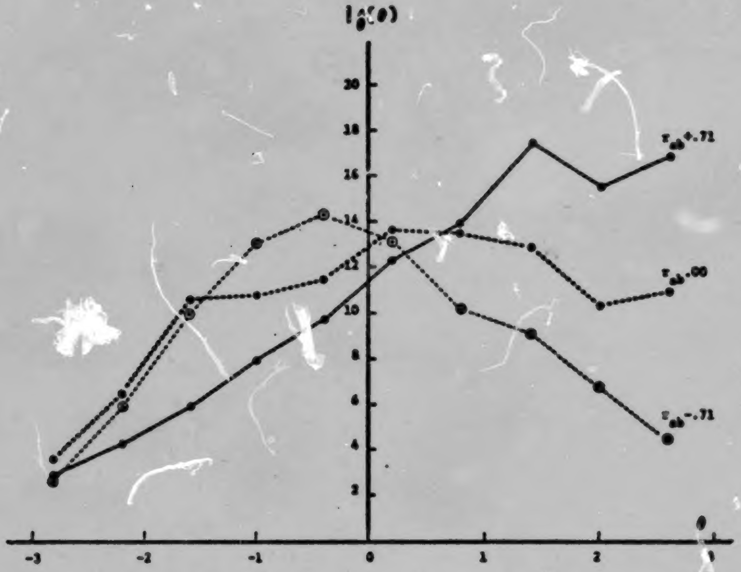


First, the information curves were calculated using a response model which may not accurately portray response tendencies of real subjects. For example, on multiple-choice tests some testees will guess. And guessing will affect the accuracy of measurement for some testees. Figure 20 shows information curves for the Bayesian adaptive test when random guessing is introduced into the response model. The curve labelled r_{ab} .00 is analagous to the data previously shown for the Bayesian strategy except that random guessing was allowed. Note that the information curve in Figure 20 is not horizontal, as it was in Figure 19. Rather, the Bayesian test provides decidedly poorer measurement for testees below mean ability (<0) than it does for those with higher abilities. Thus, comparisons of testing strategies will change as the response model changes.

A second limitation of these results is that they were based on an unrealistic item pool. First, the item pool included only 68 items;

real item pools for adaptive testing will require about 200 items per ability. Secondly, the item pool consisted of items with equal and high discriminations. If an item pool consists of items whose discriminations are correlated with their difficulties, as is usually found in real item pools, the information curves will also change shape.

Figure 20



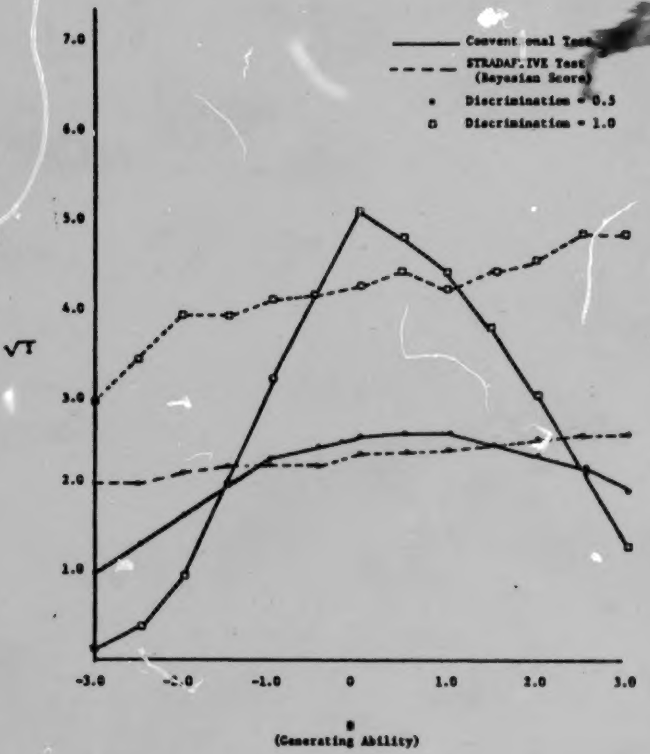
Smoothed curves of the information functions of the Bayesian sequential test under three different item pool difficulty-by-discrimination configurations.

For example, the other two information curves shown in Figure 20 were derived from the Bayesian test using an item pool in which the more discriminating items were of higher difficulty (r_{ab}+.71) and another in which the more discriminating items were less difficult (r_{ab}-.71). As Figure 20 shows, information curves under these, more realistic, item pool configurations are far from horizontal. Under both item pool configurations, the Bayesian test loses its capability of providing measurement of equal precision throughout the ability range.

Not all adaptive testing methods are as seriously affected by characteristics of the response model or the item pool as is the Bayesian strategy. Figure 21 shows information curves for the stradaptive test (and a conventional test) with guessing. With items of low discrimination (a=.5) the stradaptive information curve is still quite horizontal. For items of higher discrimination, the information curve drops somewhat for the low ability testees, but remains reasonably horizontal through most of the ability range.

Figure 21

Information Functions for 60-Item Tests



The results I've just presented are limited in yet another way. That is, all comparisons are in terms of information curves. Although information curves are a very valuable way of studying the relative utility of testing strategies, they don't tell the whole story. Testing strategies can also be compared in terms of the statistical bias in the scores they provide. Holding ability constant, bias can be defined as the difference between ability level and the average ability estimate for all testees at that ability level. If the average ability estimate is equal to the ability level, the scores are unbiased for that ability level. The greater the difference between ability and ability estimate, the greater the bias. Bias is particularly important if it differs at different ability levels or, in other words, if ability and ability estimate are curvilinearly related.

Figure 22

BIAS CURVES FOR BAYESIAN AND MAXIMUM LIKELIHOOD SCORING

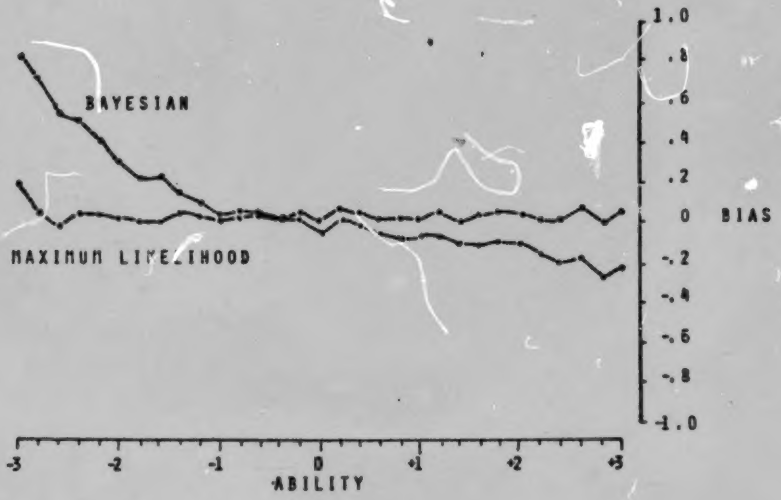
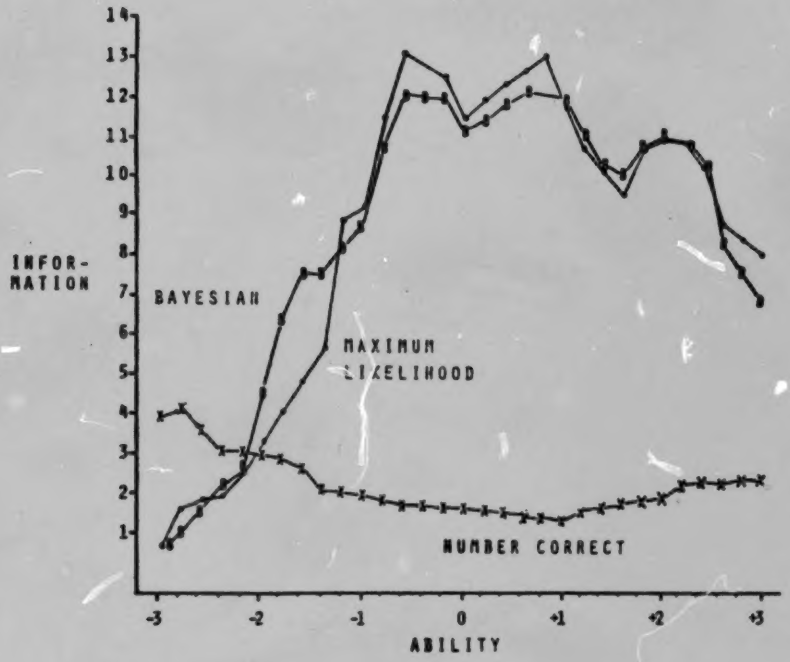


Figure 22 illustrates the bias characteristics of two methods of scoring the same adaptive test. From this we can extrapolate to the kinds of bias curves (which we haven't yet studied) which might result from two different adaptive strategies using the same scoring method. As Figure 22 shows, a Bayesian scoring technique applied to a set of data results in scores which are increasingly biased as ability deviates from the mean. Maximum likelihood scoring, applied to the same item response data, results in scores which are essentially unbiased estimators of true ability. But the information curves for the two scoring methods, shown in Figure 23, reflect very little difference between the information characteristics of the two scoring methods. Thus, different evaluative criteria (e.g., bias vs. information) can lead to different conclusions about scoring methods, in this case, or adaptive testing strategies, in general. Incidentally, Figure 23 also shows the information curve derived from number correct scoring on a Bayesian adaptive test. As can be seen, number correct score provides a very low level of information in most of the ability range. However, for very low ability testees, when random guessing is in effect, number correct score is more useful than either the Bayesian or maximum likelihood scoring methods.

Figure 23

INFORMATION CURVES FOR THREE SCORING METHODS



Some researchers in adaptive testing evaluate the "goodness" of a testing strategy in terms of the correlation between ability level and ability estimate (e.g., test score), within simulation studies. However, using this correlation as the sole evaluative criterion for comparing strategies is inappropriate, since it conceals a substantial amount of information. Figure 21 shows information functions for conventional and stradaptive tests, when both consist of items with the same discriminations. A comparison of the upper two curves in that figure show that the stradaptive test yields measurement of almost constant precision throughout the ability range, although the conventional test measures more accurately for average ability testees. The correlations of test score with ability for these data, however, are .97 for stradaptive and .95 for the conventional test. From the correlations alone, we would conclude that the stradaptive test is slightly better than the conventional test, but not dramatically so. But the information functions show considerable differences in the measurement accuracy of the two testing strategies for testees of different ability levels.

Similarly, the product-moment correlation will not reflect bias in ability estimates, as illustrated in Figure 22. Since the bias in the Bayesian score is non-linear, the correlation of ability and ability estimate will not include that non-linearity. Thus, the two scoring methods shown in Figure 22 would be evaluated similarly by the use of correlation indices, but they provide scores with quite different characteristics. And different scores will result in different decisions about people.

To summarize, we do not yet know which are the best strategies of adaptive testing. We do know, however, that adaptive tests in general have much better measurement characteristics than conventional tests, in which the same items are administered to all testees. The evaluation of adaptive testing strategies to identify those which are best will depend in part on the complex interaction of such variables as evaluative criteria, scoring methods, item pool characteristics and branching methods. We have considerable research to do on this topic, but should have some firmer answers within the next year or so.

Intra-Individual Dimensionality

As I indicated earlier, deviations from unidimensionality can result in errors in test scores. This is particularly true when summative scores are used, as they almost always are, since summation of any kind assumes one dimension underlying the responses that are summed to yield a total score. Thus, to the extent that two individuals obtain the same total score in two different ways, it can be assumed that they are not operating within the same unidimensional scale.

Some adaptive testing models permit us to begin to study the dimensionality of a particular individual's response record on an ability test. Given this capability, if we can identify different levels of intra-individual unidimensionality, resulting from the interaction of different individuals with the same "unidimensional" item pool, we can then study the consequences of deviations from unidimensionality in terms of both psychometric criteria and practical utility. One such hypothesis we can make is that scores which are unidimensionally determined should be more error-free than scores which are non-unidimensionally determined.

In ability measurement, we would expect that an individual should, in general, respond correctly to items below, or easier than, his ability level, and incorrectly to items above, or more difficult, than his ability level. If a person answers most easy items correctly and most difficult items incorrectly, we would say that he is responding consistently--that is, his response pattern seems to be influenced primarily by his position on the underlying trait continuum. However, if a person gets many easy items wrong and many difficult items correct, he is responding inconsistently, indicating that something besides the trait of interest is influencing his responses. Thus, inconsistency of this type reflects lack of unidimensionality.

In an ability test, response inconsistency may be caused by such extraneous variables (i.e., other response dimensions) as guessing, partial knowledge, or adverse psychological conditions such as test anxiety or lack of motivation to do one's best on the test. Whatever its cause, response inconsistency may reduce the reliability and/or validity of a given test score, and knowing the degree of consistency of an individual's response pattern may be important when we intend to use that score in making practical decisions.

We have operationalized the notion of response consistency in the stradaptive testing strategy. As you may recall (see Figure 15), in the stradaptive test items are organized into a series of levels or strata according to their difficulty. A correct response to an item in one stratum leads to the administration of the most discriminating item remaining in the next more difficult stratum. An incorrect response leads to the administration of the most discriminating item remaining in the next less difficult stratum.

Figure 24 shows a relatively consistent response pattern on the stradaptive test along with 10 ability scores and five consistency scores. This person entered the stradaptive test at stratum 5, based on some prior information. Stratum 5 items were too easy for him and he answered items correctly until, at item 4, he had been branched to stratum 8, which contained very difficult items. Notice that he consistently responded incorrectly to the stratum 8 items, which were too difficult for him, and correctly to the stratum 6 items, which were too easy for him. The items in stratum 7 seem most appropriate in difficulty for him, and he answered about half of them correctly and the other half incorrectly.

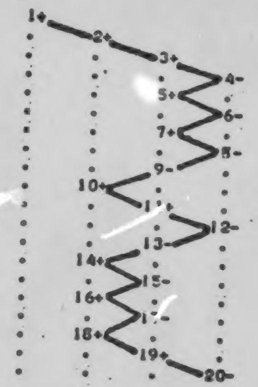
The consistency of this individual's response pattern was reflected in his relatively low consistency scores. Score 11, the standard deviation of the difficulties of the items encountered by this person, was .59. Further, in the stradaptive test, items are administered until a termination criterion is reached. Similar to the Stanford-Binet, the stradaptive test terminates when a stratum is identified at which the testee answers no items correctly, or only a chance number. The consistency of this individual's response pattern

Figure 24

Report on a Stradaptive Test for a Consistent Testee

```
                    REPORT ON STRADAPTIVE TEST
ID NUMBER:
                                        DATE TESTED:  73/07/12
-----------------------------------------------------------------

STRATUM:          (EASY)                                (DIFFICULT)
                  1     2     3     4     5     6     7     8     9
```



```
PROP.CORR:                                1.00  1.00  .56  0.00

                  TOTAL PROPORTION CORRECT=  .550
```

SCORES ON STRADAPTIVE TEST

Ability Level Scores

```
1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT=  1.40
2. DIFFICULTY OF THE N+1 TH ITEM=  1.44
3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT=  1.40
4. DIFFICULTY OF HIGHEST STRATUM
   WITH A CORRECT ANSWER=  1.33
5. DIFFICULTY OF THE N+1 TH STRATUM=  1.33
6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM=  1.33
7. INTERPOLATED STRATUM DIFFICULTY=  1.37
8. MEAN DIFFICULTY OF ALL CORRECT ITEMS=  .88
9. MEAN DIFFICULTY OF CORRECT ITEMS
   BETWEEN CEILING AND BASAL STRATA=  1.28
10. MEAN DIFFICULTY OF ITEMS CORRECT
   AT HIGHEST NON-CHANCE STRATUM=  1.28
```

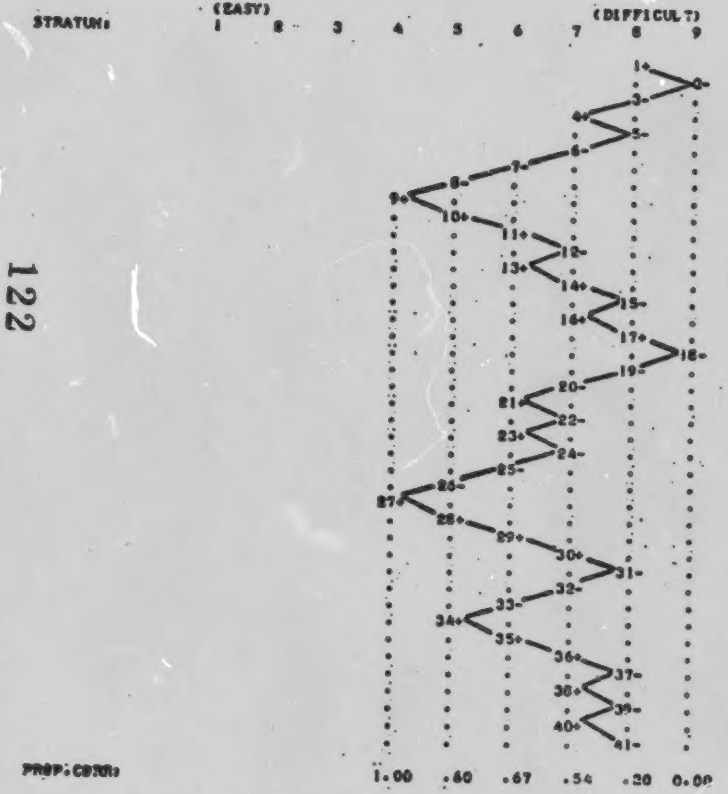
Consistency Scores

```
11. SD OF ITEM DIFFICULTIES ENCOUNTERED=  .59
12. SD OF DIFFICULTIES OF
   ITEMS ANSWERED CORRECTLY=  .46
13. SD OF DIFFICULTIES OF ITEMS ANSWERED
   CORRECTLY BETWEEN CEILING AND BASAL STRATA=  .18
14. DIFFERENCE IN DIFFICULTIES
   BETWEEN CEILING AND BASAL STRATA=  1.34
15. NUMBER OF STRATA BETWEEN
   CEILING AND BASAL STRATA=  1
```

Figure 25

Report on a Stradaptive Test for an Inconsistent Testee

```
REPORT ON STRADAPTIVE TEST
ID NUMBER:                          DATE TESTED:  73/07/02
```



```
TOTAL PROPORTION CORRECT=  .488
```

```
SCORES ON STRADAPTIVE TEST

Ability Level Scores

 1. DIFFICULTY OF MOST DIFFICULT ITEM CORRECT=  1.89
 2. DIFFICULTY OF THE N+1 TH ITEM=  1.01
 3. DIFFICULTY OF HIGHEST NON-CHANCE ITEM CORRECT=  1.53
 4. DIFFICULTY OF HIGHEST STRATUM
    WITH A CORRECT ANSWER=  2.01
 5. DIFFICULTY OF THE N+1 TH STRATUM=  1.33
 6. DIFFICULTY OF HIGHEST NON-CHANCE STRATUM=  1.33
 7. INTERPOLATED STRATUM DIFFICULTY=  1.36
 8. MEAN DIFFICULTY OF ALL CORRECT ITEMS=   .78
 9. MEAN DIFFICULTY OF CORRECT ITEMS
    BETWEEN CEILING AND BASAL STRATA=   .76
10. MEAN DIFFICULTY OF ITEMS CORRECT
    AT HIGHEST NON-CHANCE STRATUM=  1.24

Consistency Scores

11. SD OF ITEM DIFFICULTIES ENCOUNTERED=   .86
12. SD OF DIFFICULTIES OF
    ITEMS ANSWERED CORRECTLY=   .74
13. SD OF DIFFICULTIES OF ITEMS ANSWERED
    CORRECTLY BETWEEN CEILING AND BASAL STRATA=   .50
14. DIFFERENCE IN DIFFICULTIES
    BETWEEN CEILING AND BASAL STRATA=  1.64
15. NUMBER OF STRATA BETWEEN
    CEILING AND BASAL STRATA=  3
```

enabled him to meet the termination criterion after only 20 items had been administered.

Contrast this testee with the one shown in Figure 25. This person's response pattern was far less consistent and ranged over a larger number of strata. For example, this person answered some relatively easy items at stratum 5 incorrectly (note items 8 and 26) and answered some difficult items at stratum 8 correctly (items 1 and 17). By responding inconsistently, it took many more items before the termination criterion was reached, and the individual's consistency scores are higher, reflecting less consistency.

Figure 26

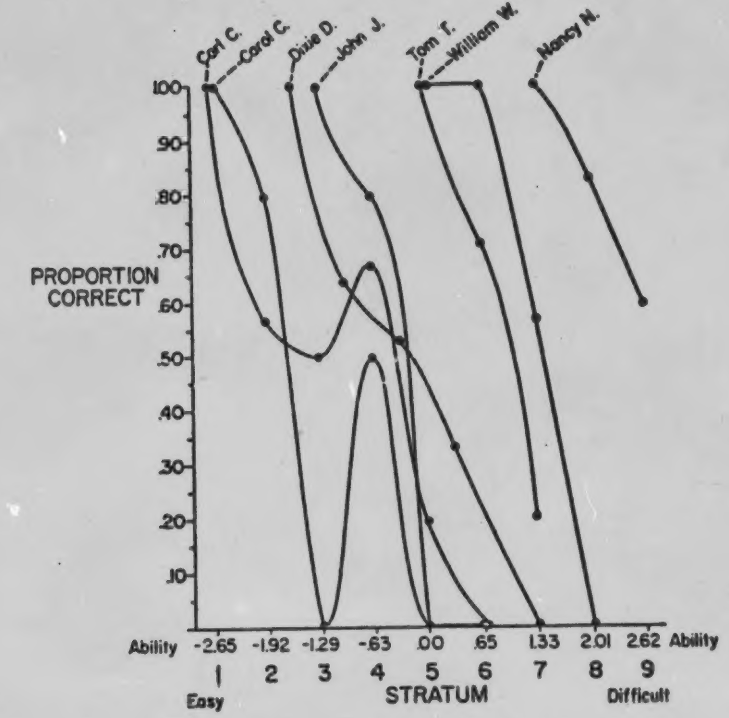


Figure 11. Proportion correct at each stratum, by individual

That consistency is related to dimensionality is illustrated in Figure 26. That figure shows a plot of proportion correct by stratum in the stradaptive test or what I have called "subject characteristic curves", for eight different testees. If the testee is responding unidimensionally, or consistently, his subject characteristic curve should show a regular decrease with increasing item difficulty--such is the case for Nancy N., William W. and Tom T., in Figure 26. For these testees, all consistency scores would be low. For the inconsistent testee, or one who is responding non-unidimensionally, proportion correct does not decrease regularly with increasing item difficulty (as is the case for Carl C. and Carol C.) or it decreases more slowly (as for Dixie D.). For these testees, inconsistency scores will be considerably higher.

We used data from live administration of the stradaptive test to study the hypothesis that the scores of individuals who are responding unidimensionally should be more error-free than those of individuals who are responding non-unidimensionally. To study this hypothesis, we used test-retest stability as an indication of score reliability, and divided a group of 200 subjects into 5 groups, according to their consistency scores on the first stradaptive test administration, in a test-retest design. Within each group, we calculated the test-retest stability of the obtained scores. Table 1 shows the results obtained for consistency score 11, the standard deviation of the difficulties of all items encountered.

As Table 1 shows, the highest test-retest stability was found in the most consistent group of examinees for all 10 ability scores. The clearest pattern is that for ability score 1, where the scores in the most consistent group had a test-retest stability of .94, and the scores in the least consistent group had a stability of .65. The stabilities in the intermediate groups decreased with decreasing consistency. Note also that the stability for the most consistent examinees on scores 8 and 9 was .98, an extremely high five-week test-retest correlation.

The possible utility of consistency scores as a moderator variable is that they might permit us to make more stable predictions for some groups of individuals (consistent testees). If these results can be replicated over longer periods of time, the consistency score might prove to be a very useful and powerful moderator variable derivable from a stradaptive testing response record. It appears to be powerful because it also moderates the test-retest reliability, but not as systematically, on the conventional test administered at the same time. Table 1 shows a test-retest reliability of .979 on the conventional test for the highly consistent group using the consistency scores derived from the stradaptive test. But consistency scores are not derivable from a conventional test so it is necessary to implement this finding within the framework of the stradaptive testing strategy.

Thus by studying the consistency or dimensionality of a set of item responses we might be able to identify individuals whose scores on a given test are more error-free. For these individuals we will

Table 1

STRADAPTIVE AND CONVENTIONAL TEST
TEST-RETEST CORRELATIONS AS A
FUNCTION OF CONSISTENCY SCORE 11
ON INITIAL TESTING

	STATUS ON CONSISTENCY SCORE 11				
	VERY HIGH	HIGH	AVERAGE	LOW	VERY LOW
MEAN CONSISTENCY SCORE	.517	.625	.706	.815	1.038
NUMBER OF TESTEES IN INTERVAL	27	30	41	43	29
STRADAPTIVE ABILITY SCORE: 1	.940	.849	.847	.768	.652
2	.875	.721	.799	.778	.751
3	.956	.813	.878	.826	.708
4	.934	.840	.846	.731	.664
5	.896	.722	.793	.756	.741
6	.950	.798	.886	.820	.704
7	.970	.844	.902	.851	.758
8	.981	.927	.915	.853	.869
9	.983	.939	.907	.899	.889
10	.951	.792	.882	.822	.718
CONVENTIONAL TEST	.979	.890	.918	.826	.878

be able to have more confidence and greater accuracy in making long-term predictions, and consequently increase our validity in the prediction of occupational criteria.

Psychological Effects

In the past, psychometricians have paid considerable attention to characteristics of tests administered to groups, for example, their reliability and validity. But we have ignored the fact that it is an individual who takes a test, not a group. Highly valid and reliable tests can be rendered useless for an individual if we do not have the cooperation of each individual or if that individual, for one reason or another, is not performing to his or her fullest capacity. For example, substantial amounts of error in the test score of an individual may result if that person's performance is hindered by high levels of test anxiety or if examinees are not motivated to do their best on each test item.

Ability tests are typically geared to the ability level of the average member of a group. Such tests will be a rather different experience for examinees of differing ability levels. The low ability individual receives a series of items which are far too difficult for him or her and may react by becoming threatened, anxious, or frustrated--the test may seem hopeless and he may simply stop trying. The high ability individual, on the other hand, receives items which are too easy for him--this person may find the task boring and unchallenging and, in a fashion similar to that of the low ability examinee, may simply stop trying to do his best. It is only for the average ability examinee that the items are likely to be sufficiently difficult to be challenging and yet not so difficult as to seem hopeless.

Adaptive testing procedures, however, tend to maintain an appropriate level of item difficulty for each individual. As a result they should keep motivation at high levels and anxiety and frustration at low levels. Or, at least, adaptive tests should equate these variables across individuals instead of, as in conventional testing procedures, allowing them to covary with ability level, which is what we are trying to measure.

Computerized test administration also allows us the capability of providing the examinee with feedback immediately after each test response as to the correctness or incorrectness of that response. Immediate knowledge of results, or feedback, may have positive motivating effects on some examinees and, therefore, they may perform at higher levels. Knowledge of results has long been considered important in the area of learning and instruction and has been built into methods of programmed and computer-assisted instruction. Further, the constructors of individually-administered intelligence tests, for example, Binet, Terman and Wechsler, stressed that some form of encouragement by the examiner was essential in keeping the examinee motivated and performing to his fullest capacity, although this encouragement was not to include knowledge of results on each test item.

Since the effects of immediate feedback on performance on objective tests of ability has been only rarely studied, we have incorporated immediate feedback into some of our research designs.

In one study[2], both a conventional test and a pyramidal adaptive test were administered by computer to a group of inner-city high school students. The group was racially mixed, consisting of both black and white students. Tests were administered such that half the group received the conventional test first, while the other half received the pyramidal test first. Within each order of test presentation, half the group received feedback and the other half did not.

---

[2] These data were analyzed by Ms. Clara DeLeon.

We analyzed the data for the conventional test only--thus, the dependent variable in this analysis was number correct on the conventional test. The design was a 2x2x2 analysis of variance. The independent variables were 1) race--black and white; 2) feedback--immediate or none; and 3) order--conventional test administered first or second in the pair.

In order to make the feedback relevant to the high school group, we had previously asked a subgroup of students from the same school to generate a set of statements which would, to them, indicate that they answered an item correctly. We used six such statements, in pseudo-random order, including "right on", "that's cool, now try this one", and "all right, how about this one". This was done on the hypothesis that feedback can have an effect only if it is meaningful or relevant to the testee.

Table 2

Mean Test Scores for Blacks and Whites on the 40-Item Test in Two Orders and With and Without Feedback

Group	Feedback		No Feedback		Total Group	
	N	Mean	N	Mean	N	Mean
Blacks--First	8	26.38	6	13.83	14	21.00
Second	7	13.86	6	14.67	13	14.23
Whites--First	15	26.07	14	30.93	29	28.41
Second	15	30.00	19	25.53	34	27.50
Blacks	15	20.53	12	14.25	27	17.74
Whites	30	28.03	33	27.82	63	27.92
First	23	26.17	20	25.80	43	26.00
Second	22	24.86	25	22.92	47	23.83
Total	45	25.53	45	24.20	90	24.87

3-Way Anova

Source of Variation	DF	Mean Square	F	Est. P
Order	1	105.76	1.36	.25
Race	1	2,013.26	25.84	<.00
Feedback	1	81.74	1.05	.31
Race x Order	1	161.54	2.07	.15
Order x Feeback	1	28.74	.37	.55
Race x Feedback	1	170.40	2.19	.14
Order x Race x Feedback	1	599.46	7.69	<.01
Error	82	77.92		

The results for the three-way analysis of variance are shown in Table 2. The only significant main effect was for race. Mean score for the blacks was 17.74 and that for the whites was 27.92, on the

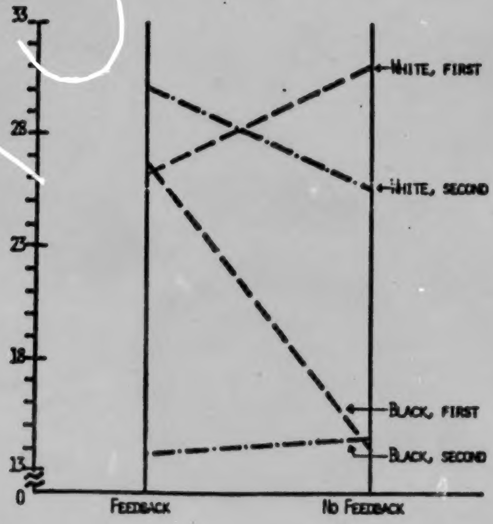
40-item test. Neither order nor feedback effects were significant, nor were any of the two-way interactions. However, the three-way order x race x feedback interaction was significant at $p<.01$.

Figure 27 shows the means for the three-way interaction. Under conditions of immediate feedback, when a conventional test was administered first, the mean of the black students (26.38) was not significantly different from the mean of the white students (26.0) who completed the conventional test under the same set of conditions.

This result implies, if it can be replicated, that race differences observed in test scores may be a function not of differences in ability but of differences in the psychological effects of the conditions of administration.

Figure 27

MEAN SCORES FOR BLACKS AND WHITES COMPLETING THE 40-ITEM CONVENTIONAL TEST FIRST AND SECOND, BY FEEDBACK CONDITION



There are some data in our results which suggest that the three-way interaction results might be due to motivational effects. In addition to analyzing test scores, we also analyzed the proportion of items skipped on the conventional test under the two experimental conditions and for the two racial groups. These results showed that blacks skipped more items than whites, in general, but when the conventional test was administered first to the black students and they received feedback, they skipped almost no items. This is also the same set of conditions under which the test scores for the blacks were not significantly different than those of the whites. This appears to be a motivational effect since when the blacks are given feedback the test becomes relevant to them; and when it becomes relevant they can answer the questions just as well as the whites.

In a second study, either a conventional test or a stradaptive test was administered with or without feedback to two groups of subjects. One group consisted of students from the College of Liberal Arts at the University of Minnesota while the other consisted of students from the University's General College. The General College group is a much less select group and has significantly lower scores on conventional ability tests. Since the tests were constructed for the higher ability Liberal Arts group, it was expected that the conventional test would be particularly inappropriate, specifically too difficult, for the General College sample.

Table 3

MAXIMUM LIKELIHOOD ABILITY ESTIMATES
FOR 40-ITEM CONVENTIONAL TEST

GROUP	FEEDBACK		NO FEEDBACK		TOTAL	
	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	-.19	57	- .52	117	- .35
GENERAL COLLEGE	28	-.88	28	-1.26	56	-1.07
TOTAL	88	-.41	85	-.76	173	- .58

TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	19.37	19.66	.001
FEEDBACK	1	5.18	5.26	.022
GROUP X FEEDBACK	1	.02	.02	.999
ERROR	169	.99		

Table 3 shows the mean maximum likelihood scores for the two groups on the conventional test according to whether feedback was or was not given. The maximum likelihood scores are in standardized units, with mean = 0.0, and s. d. = 1.0. The analysis of variance indicated a significant main effect for feedback; in both subject groups the provision of feedback resulted in significantly higher test scores. For example, in the College of Liberal Arts sample, the mean score under feedback conditions was -.19, while that under no-feedback conditions was only -.52. This difference of one-third of a standard deviation (about 3.5 raw score points) could be highly influential in a practical decision about an individual.

These results for the conventional test showed that feedback had a positive effect on test performance. But the results for the stradaptive test were quite different. Table 4 shows maximum likelihood scores on the stradaptive test under feedback and no feedback conditions. Note that in Table 4 not only is there no significant effect for feedback, but that the difference in group ability was also not statistically significant.

Table 4

ABILITY ESTIMATES FOR STRADAPTIVE TEST FOR TWO SUBJECT GROUPS WITH AND WITHOUT FEEDBACK

	FEEDBACK		NO FEEDBACK		TOTAL	
GROUP	N	MEAN	N	MEAN	N	MEAN
COLLEGE OF LIBERAL ARTS	60	-.66	62	-.62	122	-.64
GENERAL COLLEGE	28	-.96	27	-.81	55	-.89
TOTAL	88	-.76	89	-.68	177	-.72

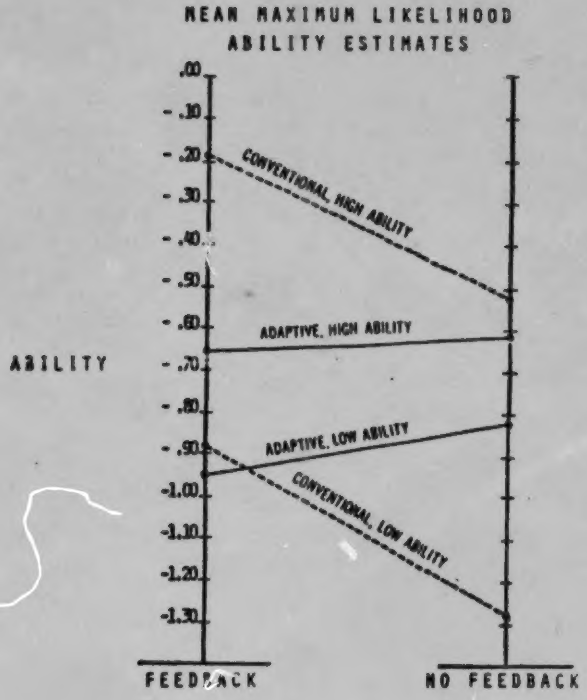
TWO-WAY ANALYSIS OF VARIANCE

SOURCE OF VARIATION	DF	MEAN SQUARE	F	EST. P
GROUP	1	2.27	1.75	.184
FEEDBACK	1	.24	.19	.999
GROUP X FEEDBACK	1	.10	.07	.999
ERROR	173	1.29		

On the surface, these results for the conventional and the stradaptive test appear to be contradictory. In the conventional test, feedback had a positive effect on test scores, and the groups differed significantly on mean ability level. On the stradaptive test, neither group differences nor feedback were significant. Figure 28 shows the means for the adaptive and conventional tests, for both groups, by feedback conditions.

If the feedback condition is interpreted as the "motivated" condition, and no feedback as "unmotivated" the apparently conflicting results can be explained. In the "low ability" (General College) group the mean

Figure 28



for the conventional test administered under feedback conditions is not significantly different from the means for the adaptive test under either condition. Or, in other words, the adaptive test itself yields scores which are intrinsically motivating to the "lower ability" testee. For the "higher ability" testee, the adaptive test scores are not significantly different from those obtained on the conventional test under unmotivated (no feedback) conditions.

The key to explaining this difference lies in the nature of the adaptive test itself. On an adaptive test--specifically, on the stradaptive test--each testee answers about 50% of the items correctly. Apparently, because of the subjective feedback the testee gets during testing, the "low ability" testee finds this "reinforcement ratio" better than what he has experienced in the past (since he is used to doing poorly on tests), and performs better even without formal feedback. The "high ability" testee, on the other hand, is used to getting a large proportion of items correct on a conventional test. But he finds the adaptive test much more difficult than he is used to, and may experience some of the frustration that the typical low ability testee usually encounters. The fact that the mean ability estimates for the high ability group were not significantly different from those of the low ability group on the adaptive test, suggests that the adaptive test reduces error variance for the low ability testees which artificially depresses their test scores.

These results are obviously not conclusive and replications and further studies are certainly necessary. But given the current furor over test fairness and bias, it seems that we should pursue further the effects of various conditions of test administration upon performance, particularly for "low ability" testees, whose abilities might not be so low after all. Adaptive testing and immediate knowledge of results may be able to provide testing conditions more conducive to each individual's capability to demonstrate his/her fullest capacities in test performance. And, since computerized adaptive trait measurement, can provide us with important additional information of a variety of types, as well as providing more precise measurement throughout the ability range, it has promise of supplanting the paper and pencil tests which have dominated psychological testing for the last 50 years.

DEVELOPMENT OF QUALIFICATION STANDARDS AND EVALUATION PROCEDURES -A PRACTICAL APPROACH

Alex L. Szaszy
Defense Language Institute

BACKGROUND

The Defense Language Institute regularly conducts training in as many as thirty different foreign languages. Faculty turnover (approximately 20%) requires the almost continual hiring of new personnel, only a few of whom have had previous experience with the unique characteristics of the DLI mission, methods and student body. The difficulties of faculty hiring are further compounded by the fact that most applicants received much or all of their education and teaching experience outside the U.S., which makes interpretation and comparison of academic records particularly difficult, and also raises huge problems in determining and comparing the quality and relevance of past teaching and other experience. Similar problems exist in promoting FL instructors from entry grade (GS-7) to journeymen (GS-9) and to supervisory and Department Chairman positions. Since all these positions are in the Excepted Service, the CSC qualification and classification standards are inapplicable.

In April 1973, the Defense Language Institute forwarded a draft of DLI Qualification Standards for Foreign Language Instructional Personnel to HQ, Department of the Army, Deputy Chief of Staff for Personnel, Office of Civilian Personnel (DAPE-CP) for review and comment. DAPE-CPP replied to DLI stating that the standards need to be reevaluated in terms of job analysis and job-related criteria, and be in conformance with Equal Employment Opportunity requirements and Civil Service Commission policy.

Furthermore, since the CSC had determined that some rating procedures being used in DA violate CSC policy and regulations concerning the use of education in candidate evaluation, DAPE-CPE issued specific guidelines in this matter. These guidelines stated that all evaluation criteria, methods and instruments must be based on a thorough job analysis and must be relevant to the skills, knowledges, abilities and personal characteristics important to successful performance in the job(s) to be filled. Education in general, a particular curriculum, or a specific course could be considered in the evaluation process only if it is likely to indicate that the individual was exposed to and had the opportunity to develop one or more of the skills and knowledges identified through job analysis as important to superior performance in the job(s) to be filled. The total educational background along with other available information, such as work history, training background, self-development activities, supervisory appraisals of performance or assessment of potential, and awards is reviewed to determine to what degree the individual possesses the essential skills and knowledges. This does not mean that educational background is no longer an important consideration when evaluating candidates for placement/promotion. The educational background is part of the "total information package" to be reviewed to determine , to a high degree of certainty, whether or not the candidate possesses the kinds and levels of skills and knowledges important to job performance.

FOLLOW-THROUGH

In May 1973, DLI established a Qualification Standards Committee to develop Qualification Standards for instructional personnel (GS 07-12) in consonance with DA and CSC procedures and policy. The Committee devised an evaluative survey questionnaire (Job Element Inventory Survey) in order to conduct a thorough job analysis. Based on the results of the job survey, the major job-related duties and requisite skills, knowledges and abilities (SKA) were identified for each instructional position (GS 7-12, Excepted Service), and the Qualification Standards, developed by the Committee, were approved by the DA. The DLI Committee also developed procedures and instruments to evaluate the candidates' teaching, supervisory and managerial skills, knowledges, abilities and personal characteristics (SKAPs), including a rating and ranking system to identify the highly qualified applicants/candidates. The system became operational in November 1973.

TABLE 1

While reviewing the evaluation procedures during the preliminary assessment process, it was noted that one of the main problems in the assessment of candidates was the fact that the SKAP items listed in the Qualification Standards were too numerous to permit their practical use. It was also evidenced that the evaluation system didn't discriminate at the GS-7 and GS-9 level. For example, ninety per cent of the applicants for the GS-7 instructor position were rated as highly qualified.

TABLE 2

Because of this problem, DLI decided to form a Task Force to revise the Qualification Standards by reducing the SKA items into manageable proportions to achieve simplicity and operational clarity with respect to the behaviors to be observed and rated. This effort was limited to the current revision of the GS-9 standards since the GS-9 level teaching SKAs are common to the positions at the GS-7 through GS-12 level. In addition, the decision in a case now pending before the Supreme Court concerning current CSC restrictions on non-resident aliens in the Competitive Service may impact on DLI's Excepted Service positions. If conversion to the Competitive Service takes place the only placement actions internal to the DLI Civilian Personnel Office will be the promotion of permanent GS-7 foreign language instructors to GS-9 (as part of the Merit Placement and Promotion Plan), since eligibility for GS-7 temporary and permanent appointments would be determined by the CSC and initial hiring at and promotion to the GS-11 and above levels would become part of the DA Education and Training Career Program.

THE REVISION PROCESS

A proposal, subject: Proposed Procedures for Advising DLIs Revision of Hiring and Promotion Practices, submitted by Dr. Jack H. Hiller, Senior Project Director, Army Research Institute (ARI), R&D Field Activity, Fort Ord, California dated 12 July 1974, **Appendix A**[1]**, was used as a mode.** The proposed steps for revising the effectiveness and the evaluation technique were generally followed and accomplished by the Task Force with ARI support on a Technical Advisory Services (TAS) basis. Since the stages, listed in the proposal, included several steps, such as refinement of the GS-9 Qualification Standards, rating of the revised elements (SKAs) on the basis of importance to teaching (impact on student learning) and discrimination capability, DLI decided that this plan be accomplished in two phases. In Phase One, revision of the Qualification Standards, and in Phase Two, the exploration of the usefulness, validity and feasibility of several measurement techniques for use in translating the SKAP items into measurable terms.

Revision of the Qualification Standards

One of the most critical tasks in the revision process was to re-write and/or revise each SKA (skill, knowledge and ability) element in the Qualification Standards, to achieve simplicity and operational clarity. Each item was reviewed and revised for clarity, elimination of redundancy and possible regrouping. The main objectives during this process were: (1) to reduce the number of SKA items into manageable proportions, and (2) to revise and/or rewrite each item in such a way that the particular skill or ability could be demonstrated and measured. An item-by-item check followed this exercise to insure that the items were as clear as possible and could be used for rating. SKA items were then grouped into two main categories: (1) classroom instruction - teaching related, and (2) instructional material development related items, since the GS-9 instructor is required to perform both duties.

Selection of Raters

A group of highly experienced DLI instructional/staff personnel were identified (representing all GS grade levels and languages), to participate in the rating of the SKA items.

---

[1] Not included in these proceedings. Persons desiring copies of Appendix A should contact the author.

TABLE 3

In addition to the rating group, the National Federation of Federal Employees (NFFE), Local 1263, The National Association of Supervisors, Federal Government (NAS), Chapter 107, the Academic Advisor, and the Office of Research and Development were also asked to participate in the rating as an independent group. Although major decisions were made on the basis of the ratings completed by the rating group (44 Instructional/staff personnel), ratings of the independent group (Union, Management) were also considered.

Development of Rating Scale

The criterion for the inclusion of any SKA item in the Qualification Standards was that each item should be judged by experts as follows: (1) the behavior (SKA) occurs frequently in the classroom and, therefore, can be observed (measured) in most classroom situations, (2) the item is important, i.e., has a relatively high impact on student learning, and (3) the item has the capability to distinguish on a statistically significant basis between satisfactory and superior performance. A 5-point rating scale was developed for the judges to rate each SKA item in the following dimensions:

a. Relative frequency of use or need. The judges were asked to rate each SKA item on the basis of the frequency of use or application of a certain item by the instructor in order to be successful in the conduct of his/her classroom instruction, considering a typical DLI instructor who is teaching all phases of a DLI course. Items were rated as: one of the most frequently used (5 points), highly frequently used (4 points), frequently used (3 points), sometimes used (2 points) and seldom used (1 point).

b. Criticality. The judges were asked to rate each SKA item on the basis of its importance to foreign language teaching, or on the basis of the degree of impact each SKA has on student learning. Items were rated as extremely critical (5 points), very important (4 points), important (3 points), have some importance (2 points), or have little importance (1 point).

c. Discrimination capability. Judges were asked to rate each SKA item on the basis of how well the particular item can discriminate among candidates. Items were rated as very discriminating (5 points), highly discriminating (4 points), moderately discriminating (3 points), low discrimination capability (2 points), and no discrimination capability (1 point). A copy of the rating scales used by the raters is attached as Appendix B[2].

---

[2] **Not included in these proceedings. Persons desiring copies of Appendix B should contact the author.**

Data Analysis

Ratings of the 44 judges (rating group) were tabulated for each item and the mean scores were computed in each dimension (frequency, criticality and discrimination). Additionally, since the Local Union represents the majority of the faculty, different weights were assigned in computing one mean score for the Union/Management group. Thus, within the independent group, the Union and Management were given equal weight in the ratings of each SKA item.

Since the main purpose of the rating was to reduce the element list, cut-off points were set. These cut-off points were based on a complex judgment and considered the absolute values of the ratings and the number of items that would remain in the Qualification Standards. Generally, a mean score of 3.6 on the discrimination scale was used as the cut-off point. The rationale behind this was that items which have lower discrimination power (mean score less than 3.6) may not identify the highly qualified candidates as well as items with higher discrimination capability (mean score of 3.6 or higher). The interpretation of the 3.6 mean rating indicates that the item will discriminate not necessarily "highly" (a 4.0 point value on the rating scale), but certainly more than "moderately" (3.0 point value). Moreover, the evaluation procedure will be simplified, or more eficient, if the items selected for the rating instruments have more than "moderate" discrimination power.

When the mean rating of the selected raters was significantly higher than the rating of the other group (NFFE, NAS, AA and R&D), the item was considered acceptable to remain on the list even if the mean rating of the other group was below 3.6 (ranging from 3.59 to 3.0). However, in this case, the item had to have a mean rating of 4.0 or higher by the selected raters. After the cut-off points were set and items were selected on the basis of the mean ratings on the discrimination scale, the items were ranked according to the mean values on the combined frequency/criticality scale.

TABLE 4

The revision and the rating process resulted in the drastic reduction of the number of elements (SKAs) in the Qualification Standards but more importantly in the identification of those SKAs which are the most critical on the job and have high discrimination power. Thus, specific measurement techniques can be developed for those 14 skill, 15 ability and 8 knowledge items which are being included in the revised GS-9 Qualification Standards.

TABLE 5

IDENTIFICATION OF MEASUREMENT TECHNIQUES

The next step is that two major types of tests, i.e., performance tests to measure skills, and paper/pencil tests to measure knowledge, will be developed. Abilities can be measured using performance tests and also by paper/pencil tests. Given the need to tie certain measures to specific languages, it would be impractical to attempt to produce all of the required tests, and it will be necessary to develop prototype measures and guidelines for preparing the tests.

Development of a performance test requires the following steps:

- Describing test situations that specify the conditions under which the performance is to be demonstrated (or the target skills are to be measured).

- Describing in detail the activity to be performed by the candidate (the critical criterion performance that will be measured).

- Preparation of rating scales that will be used to measure the target skill characteristics.

- Conducting experiments, using both experienced and beginning instructors, so that the selection of tasks to be performed, the clarity of the directions which describe the test, and the equipment, materials, and time limits can be verified.

- Development of a scoring system and revision of the test using the information provided by the experiment.

- Development of a detailed test manual providing a rating scale (check list) for each critical aspect in the test operation.

- Training of raters (supervisors) to insure that the measurement system produces reliable and valid measurements.

- Scoring the resulting data and developing norms or cut-off scores for acceptable performance.

The knowledge tests may take several forms. Standard multiple choice tests may be used to test knowledge of an academic nature, such as knowledge of pedagogical principles, foreign language teaching methodology, linguistics, etc. Knowledge of methodology may also be tested by having the instructor (candidate) rate the performance of a filmed or videotaped model. Knowledge related to auditory discrimination or pronunciation may be tested with a tape recording. All tests must be subjected to empirical testing to insure reliability and validity. It is anticipated that the evaluation system will be more objective and simplified.

TABLE 6

SUMMARY

The primary purpose of personnel selection and placement is to predict job behavior. The central role of job analysis information (and as a result, the identification of skills, knowledges and abilities necessary for successful job performance) is to provide the basic clues for choosing potentially useful predictor measures and for developing the measures of job behavior DLI desires to predict.

It is important, therefore, that the identified skills, knowledges and abilities in the Qualification Standards should not only be job-related but should also predict job behavior. The ultimate intent of the DLI effort was to provide a basis for developing reliable and technically adequate selection criteria and measurement devices whereby the most promising individuals can be selected from a larger group of candidates for the position of a foreign language instructor at the Defense Language Institute.

TABLE 1

RATING PROCEDURE FOR DLI INSTRUCTIONAL PERSONNEL, EXCEPTED SERVICE, GS 1710-07/12

POSITION AND GRADE	RATING FACTORS		MINIMUM CRITERIA	MODE OF EVALUATION	RATING FORMULA
Instructor GS-07	Foreign Language	Speaking	Level-4	Interview	(FL speaking)+ (FL writing)+ (EL speaking)+ (EL writing)+ (Teaching SKAs X 4) = Total Points
		Writing		Essay	
	English Language	Speaking	Level-1	Interview	
		Writing		Essay	
	Teaching SKAs (Skills, Knowledges, Abilities)		Overall B. S. Rating	Classroom Obs Interview	
Instructor GS-09	Foreign Language	Speaking	Level-4	Interview	(FL speaking)+ (FL writing)+ (EL speaking)+ (EL writing)+ (Teaching SKAs X 4) = Total Points
		Writing		Essay	
	English Language	Speaking	Level-2	Interview	
		Writing		Essay	
	Teaching SKAs		Overall B. S. Rating	Classroom Obs Interview	
Class Supervisor GS-11	Foreign Language	Speaking	Level-4	Interview	(FL speaking)+ (FL writing)+ (EL speaking)+ (EL writing)+ (Teaching SKAs X 4)+ (Supervisory SKAs X 4) = Total Points
		Writing		Essay	
	English Language	Speaking	Level-3	Interview	
		Writing		Essay	
	Teaching SKAs		Overall B. S. Rating	Classroom Obs Interview	
	Supervisory SKAs		" "	Interview	
Department Chairman GS-12	Foreign Language	Speaking	Level-4	Interview	(FL speaking)+ (FL writing)+ (EL speaking)+ (EL writing)+ (Teaching SKAs X 4)+ (Supervisory SKAs X 4)+ (Management SKAs X 4) = Total Points
		Writing		Essay	
	English Language	Speaking	Level-3	Interview	
		Writing		Essay	
	Teaching SKAs		Overall "Barely Satisfactory" Rating	Classroom Obs Interview	
	Supervisory SKAs			Interview	
	Management SKAs				

TABLE 2

NUMBER OF ITEMS (JOB ELEMENTS; DUTIES; AND SKILLS, KNOWLEDGES, ABILITIES - SKAs) LISTED IN THE QUALIFICATION STANDARDS

Item		Positions							
		GS-7		GS-9		GS-11		GS-12	
Job Elements		3		4		10		11	
Duties		23		26		44		51	
SKA	Skills	38	106	58	170	73	248	87	320
	Knowledges	22		39		73		104	
	Abilities	46		73		102		129	

TABLE 3

COMPOSITION OF RATING GROUP (JUDGES)

Language Group	Grade Level		Total
	GS-11/12/13	GS-9	
Asian	7	6	13
Romance/ Germanic	5	4	9
Slavic	8	6	14
Others	4	4	8
Total	24	20	44

WEIGHTS ASSIGNED FOR THE UNION/MANAGEMENT RATING GROUP

Organization	Weight (points)
National Federation of Federal Employees, Local 1263	5
National Association of Supervisors, Chapter 107	2
Academic Advisor	1
Research and Development	2

TABLE 4

PART I, CLASSROOM RELATED ELEMENTS

Item No	Mean Ratings by Judges			Ratings on Discrimination Scale		Items Selected Based On Discrimination Ratings	Ranks Based On Frequency/ Criticality Mean Rating
	Frequency Scale	Criticality Scale	Mean Frequency/ Criticality	Judges	Other Groups NFTE,NAS,AA,R&D		
					Weight		
					5 2 1 2		
1	4.30	4.50	4.40	4.23	3.9	X	7
2	3.95	4.20	4.08	4.02	4.1	X	19
3	2.84	3.07	2.96	2.84	2.9		
4	3.18	3.36	3.27	3.30	3.6		
5	3.77	3.93	3.85	3.82	4.3	X	23
6	4.64	4.59	4.62	4.16	4.2	X	2
7	3.25	3.68	3.47	3.55	4.1		
8	4.02	4.30	4.16	4.00	4.1	X	14
9	3.73	3.93	3.83	3.52	4.4		
10	4.39	4.43	4.41	4.14	3.9	X	6
11	3.18	3.39	3.29	3.36	3.2		
12	3.64	3.50	3.57	2.98	3.4		
13	3.30	3.55	3.43	3.61	4.2	X	26
14	3.61	3.89	3.75	3.86	3.8	X	24
15	3.45	3.93	3.69	3.77	3.4		
16	2.14	2.98	2.56	3.09	2.4		
17	3.18	3,50	3.34	3.20	3.3		
18	3.59	3.73	3.66	3.36	2.7		
19	4.00	4.43	4.22	4.02	3.4	X	9
20	4.16	4.18	4.17	3.95	3.6	X	13
21	4.11	4.05	4.08	3.91	2.9		
22	4.61	4.68	4.65	4.11	4.2	X	1
23	3.70	4.05	3.88	4.02	3.0	X	22
24	3.61	3.82	3.72	3.68	3.9	X	25
25	4.25	4.11	4.18	3.86	3.7	X	11

TABLE 4 (Cont.)
PART I. CLASSROOM RELATED ELEMENTS (Cont.)

Item	Mean Ratings by Judges			Ratings on Discrimination Scale		Items Selected Based On Discrimination Ratings	Ranks Based On Frequency/ Criticality Mean Rating
No	Frequency Scale	Criticality Scale	Mean Frequency/ Criticality	Judges	Other Groups NFFE,NAS,AA,R&D Weight 5 2 1 2		
26	2.25	3.11	2.68	3.34	3.0		
27	4.07	4.12	4.10	3.96	4.2	X	18
28	3.80	3.86	3.83	3.71	3.1		
29	3.88	4.04	3.96	3.89	4.0	X	21
30	4.25	4.18	4.22	3.98	4.8	X	10
31	4.14	4.09	4.12	4.07	3.4	X	16
32	4.18	4.16	4.17	4.07	4.0	X	12
33	3.96	4.02	3.99	3.60	3.9	X	20
34	3.32	3.77	3.55	3.98	2.9		
35	3.09	3.34	3.22	3.20	2.6		
36	3.07	3.30	3.19	3.45	3.8		
37	3.62	3.71	3.67	3.50	4.4		
38	3.36	3.39	3.38	3.96	4.9	X	27
39	3.00	3.13	3.07	2.93	3.8		
40	3.52	3.61	3.57	3.29	3.2		
41	2.82	3.11	2.97	3.28	3.1		
42	4.57	4.41	4.49	4.00	4.0	X	5
43	4.48	4.52	4.50	4.41	3.8	X	4
44	4.34	4.41	4.38	4.39	3.6	X	8
45	4.48	3.98	4.23	3.32	3.5		
46	4.52	4.52	4.52	4.23	4.1	X	3
47	4.09	4.16	4.13	4.11	3.3	X	15
48	4.02	4.18	4.10	4.09	3.6	X	17

TABLE 4 (cont.)
PART II. DEVELOPMENT RELATED ELEMENTS.

Item No	Mean Ratings by Judges on Criticality Scale	Ratings on Discrimination Scale		Items Selected Based On Discrimination Ratings	Ranks Based on Ratings or Criticality Scale
		Judges	Other Groups NFFE,NAS,AA,R&D Weight 5 2 1 2		
1	4.55	4.32	4.6	X	3
2	3.73	3.61	2.9		
3	4.16	4.14	3.5	X	9
4	3.75	3.60	3.2		
5	3.70	3.55	3.8		
6	4.30	4.11	3.5	X	7
7	4.61	4.54	3.3	X	2
8	4.39	4.21	4.3	X	5
9	3.20	3.30	2.8		
10	4.32	4.41	3.0	X	6
11	3.60	3.48	3.3		
12	4.20	4.00	3.3	X	8
13	3.77	3.75	4.1		
14	3.88	3.79	3.7		
15	4.14	3.91	3.8		
16	4.21	3.91	3.3		
17	4.25	4.13	2.5		
18	4.64	4.32	4.4	X	1
19	4.00	4.07	3.1	X	10
20	4.46	4.16	3.4	X	4
21	4.11	3.91	3.7		

TABLE 5

COMPARISON OF ITEMS IN THE GS-9 QUALIFICATION STANDARDS

Item		Number of Items		
		Original Qualification Standards	After Revision (before rating)	Finalized Qualification Standards
Job Elements		4	NA	3
Duties		26	NA	16
SKA Items	Skill	58	20	14
	Knowledge	39	14	8
	Ability	73	33	15

TABLE 6

COMPARISON OF RATING PROCEDURES FOR GS-9 INSTRUCTIONAL PERSONNEL (EXCEPTED SERVICE)

CURRENT SYSTEM

RATING FACTORS		MINIMUM CRITERIA	MODE OF EVALUATION	RATING FORMULA
Foreign Language	Speaking	Level-4	Interview	(FL speaking points)+ (FL writing points)+ (EL speaking points)+ (EL writing points)+ (Teaching SKAs points X 4) = Total Points
	Writing		Essay	
English Language	Speaking	Level-2	Interview	
	Writing		Essay	
Teaching SKAs (skill, knowledge, ability)		Overall B. S. Rating	Classroom Observation	
			Panel Interview	
Educational background		N/A	Review of background info	

PROPOSED SYSTEM

RATING FACTORS		MINIMUM CRITERIA	MODE OF EVALUATION	RATING FORMULA
Foreign language	Speaking	Level-4	Interview	(FL speaking points)+ (FL writing points)+ (ECL points)+ (Performance Test points X 4)+ (Knowledge Test points X 2) = Total Points
	Writing		Essay	
English language		70 ECL score	English Comprehension Level Test (paper-pencil)	
Teaching SKAs (skill, knowledge, ability)		Overall B. S. Rating	Performance Test (observ.)	
			Knowledge Test (paper-pencil)	
Educational background		N/A	Review of background info	

DELPHI APPLICATIONS TO JOB TASK ANALYSIS

Mr. John F. Kaczmarski

United States Army Infantry School

(SLIDE 1)

Any systematic approach to training requires a rigorous and definitive description of the end-product of that training. In an attempt to produce such a description, training managers have constructed job task inventories that have generally been reflected in the instruction. However, to be most effective as a training and management tool, a job task inventory must reflect the total package of tasks an individual performs in executing his job; not merely those tasks in which he was formally trained.

By taking a "whole man" approach to task analysis, it was deemed possible to produce a prioritized task inventory that may be used as:

(SLIDE 2)

A management tool for the allocation of resources.

A training tool for the production of behavioral objectives, instruction and tests.

A baseline evaluation tool to determine the overall success of the program.

These three were accepted as the criteria for a useful task inventory. In attempting to produce a task inventory that would meet these criteria, recent methods being used in task analysis were reviewed. Current literature revealed that while most training models provide for the equivalent of task analysis, specific descriptions of methodologies are at a premium. Generally, the methods of tasks analysis fall into 5 categories:

(SLIDE 3)

Interviews
Questionnaires
Simulations
Brain Storming
Direct Observation

Without exception, each of these techniques operates by soliciting expert opinion or expert observations of the perceived tasks of a job.

While each of these techniques has specific advantages, all were rejected in favor of a technique that appears to enjoy most of the same advantages and few of the disadvantages of the previously listed techniques. That is the

(SLIDE 4)

DELPHI TECHNIQUE

When the Delphi Technique was first developed at the Rand Corporation, in the 1960's, it was designed to be used as a political and economic forecasting tool. The Delphi achieved a large degree of acceptability with the highest levels of the economic and political communities. Over the years the Delphi Technique has generalized from its original purpose to many other fields. There have been recent applications in such diverse areas as: defining corporate goals, developing military strategy, determining the educational needs of communities, and refining the direction of teacher education programs. During this transition period, the Delphi was operantly defined as a method of eliciting expert opinion and achieving consensus on that opinion. While there are adherents to the use of Delphi as a forecasting tool, most of the recent applications have been to define courses of action where large bodies of opinion and little hard data exist. In short, the mysticism that surrounded the Delphi Technique at its inception has dissolved in favor of a more nuts and bolts approach to determining what is, rather than what may be. This is the application in which Delphi may be most fruitfully viewed as a method of task analysis.

In general terms, the Delphi Technique operates through a panel of subject matter experts. The experts work in isolation using a series of questionnaires designed to extract their individual expertise and mold it into a group opinion. Most of the current literature indicates that modifications have been made to the technique. Most often, these modifications have been in the size of the panel of experts, the number of rounds of questionnaires required to achieve consensus, and the method of providing feedback to the panel. They do not represent substantive changes.

The Delphi Technique was recently used on an experimental basis to develop a job task inventory for Pathfinder training at the

US Army Infantry School. A Pathfinder provides navigational assistance to Army aviation; however, a few students are also trained from the Navy, Air Force and Marine Corps, as well as Allied officers.

The development of the Pathfinder job task inventory was phased into two separate operations; development of the task list, and prioritization of the task list.

PHASE I

In developing the inventory, the initial step was to identify and analyze the duty positions and organizations to which the graduates would be assigned. This was done to determine if there were major fluctuations in the tasks the US Army graduates must perform. It was determined that duty position, organization and military rank did not significantly influence the tasks that a Pathfinder performs. From this initial step, it became apparent that, in terms of training needs and job tasks, prospective Pathfinders were a homogeneous group, and that a single task inventory would suffice for the officers and enlisted men of Army organizations.

THE PANEL

Developing a panel of experts to generate the task inventory posed a more difficult problem. Despite the apparent homogeneity of the tasks of Pathfinders, it was decided that the panel should represent a stratified sample of the various Pathfinder units and the military grades within those units.

Additional requirements were that all panelists must be formally qualified and possess recent experience as Pathfinders. Further, each panelist must be perceived by his peers as an expert in his field. All of these requirements and precautions were taken because the task inventory would be elicited from the combined experience and expertise of the panelists. Bias among the panel members would result in a biased task inventory.

A series of brain storming sessions was held with USAIS resident experts to develop a list of nominees. Contact with these nominees generated additional nominations. Subsequently, the list was narrowed to 10 individuals who met the criteria for the panel and who were willing and able to devote their time. (Limiting the panel to 10 participants was based on a decision that 10 was the largest number of panelists that could be readily managed by the controller.

After the composition of the panel had been finalized, each participant received a memorandum confirming their membership on the panel. Additionally, the procedures of the Delphi Technique were re-explained and special emphasis was placed on the need to work in isolation. The participants were allowed and encouraged to use any references they wished in preparing their input, but cautioned that what was desired was their expertise, not a summary of publications.

It was estimated that it would take 5-6 rounds to generate, prioritize and achieve consensus on the task inventory. In fact, it was achieved in four rounds and two mini-rounds. For the purpose of writing this task inventory, consensus was operantly defined as agreement by 80% of the panel members.

FIRST ROUND

Panelists received a memo briefly reviewing the procedures, and asking them to "list those tasks a Pathfinder must be able to perform in successfully executing his missions." The request was deliberately couched in as non-directive, open-ended terms as possible. The attempt was to elicit as wide a range of responses as possible, and this was emphasized to the members. The panelists were issued forms that were designed to induce short, 4-6 word task statements. Examples of task statements from other inventories were furnished the panelists.

The panel generated a total of 296 task statements with a range of 36, a median of 30 and a mode of 38. This great range was indicative of the difference in scope of the task statements generated. At this point two non-participating subject matter experts were consulted. They were of great assistance in collating and translating the data into single forms. This collation of data produced 91 task statements. Unanimity was achieved on 2 tasks and consensus achieved on an additional 5 task statements.

It should be noted that 41% of the tasks could be characterized as belonging to the cognitive domain.

(SLIDE 5)

No tasks were identified from the affective domain, and 59% of the tasks could be categorized into the Psycomotor Domain.

While no tasks were identified from the affective domain, some of the tasks, such as making a H.A.L.O. insertion (pinpoint free-fall parachute operations), imply that certain attitudes are a prerequisite to a task.

The breakout of the frequency of which each task was proposed by a panelist was as follows:

(SLIDE 6)

SECOND ROUND

After the data was collated and computed, each task statement was listed on task inventory forms. Immediately to the right of each task, a number was shown indicating the percentage of panelists nominating each task. The purpose of this data was to provide the panel member information on the perception of his co-panelists. A copy of the inventory was furnished the panelists with the same cautions as in the first round. The panelists were asked to indicate agreement or disagreement with each task. If they had particularly strong feelings on the pros or cons of any task, they were invited to make written comment in the space provided. Additionally, panelists were asked to record any new tasks that had been generated since the first round.

Eight pencil changes were made to indicate a need to rephrase a task.

The second round generated no new task statements.

The marked increase in consensus among the panelists on task statements may be seen here.

(SLIDE 7)

As you may see, consensus was achieved on all but 4 tasks. The four tasks were disposed of in a subsequent mini-round.

This dramatic shift emphasizes the ability of the Delphi Technique in inducing consensus.

The first phase of writing the task inventory was now essentially complete. A listing of the broad job tasks of Pathfinders was produced. This is where most task inventories stop, no matter the technique used. Essentially what has been generated is merely a listing of tasks. The potency of this task inventory to meet the criteria stipulated is severely handicapped by a lack of any prioritization of tasks; by prioritizing these tasks, the degree to which the criterion may be met is greatly increased.

The second phase utilized the same techniques and panel to prioritize the inventory into three categories:

(SLIDE 8)

PHASE II

1. *Prerequisite Tasks* - Job tasks in which a combat arms soldier would normally be proficient prior to arrival at USAIS for Pathfinder training.

2. *Resident Training Tasks* - Job tasks placed in this category should meet one or more of the following standards:

a. This task is critical to achievement of Pathfinder missions.

b. The facilities, equipment or expertise necessary to teach this task are only available at The Infantry School.

c. Standardization of the method or degree to which the student achieves the task is critical to Pathfinder missions.

3. *On The Job Training* (OJT) - Tasks which can reasonably be mastered after completion of training at The Infantry School and assignment to a Pathfinder position. These tasks are not critical to the achievement of Pathfinder missions.

Another force behind this attempt to prioritize the instruction was recognition of the building block nature of most learning. Success in this categorization would allow for a logical progression of instruction and mesh resident and non-resident training. An additional benefit accrued is that it assures achievement of basic skills prior to resident training and allows concentration of school resources on those tasks which can *only* be taught in residence.

THIRD ROUND

The results of this initial attempt at prioritization were exceptional. Consensus was achieved on 79% of all tasks. The distribution may be seen here.

(SLIDE 9)

An analysis of the 21% of the tasks in which consensus was not achieved indicated that the difference of expert opinion was generally over whether a task should be a prerequisite or OJT task, rather than a split over resident vs non-resident training.

Neither the subject matter nor the scope of the task appeared to be germaine to the conflict.

FOURTH ROUND

The agreed upon tasks were removed from further consideration and the remaining 18 tasks were resubmitted to the panel for their reconsideration. The panel achieved consensus on 12 of the 18 tasks. The remaining 6 tasks were disposed of by an additional mini-round.

The results of the third and fourth round categorized the tasks as follows:

(SLIDE 10)

The major change produced by this last round was the increase in the OJT tasks by 10%. Prerequisite and formal training tasks were both increased about 5%.

When finalized, the task list was composed of 87 tasks divided into three categories:

Prerequisite - 16 Tasks - 18%
Resident - 60 Tasks - 69%
OJT - 11 Tasks - 13%

It should be noted that the majority of the cognitive tasks were grouped in the formal training category. Tasks categorized into the prerequisite category were primarily from the Psycomotor Domain.

It is noteworthy that up until this point only 69 manhours had been expended by panelists, controller and consultants in preparing the task inventory.

The prioritized task inventory may be used to meet the previously listed criteria.

(SLIDE 11)

As a management tool for the allocation of resources, the task inventory allows the resident training manager to:

1. Prevent resources from being allocated to teach tasks that have already been mastered by students or can easily be mastered after graduation.

2. Select candidates with skills necessary to successful completion of training.

3. Organize his instructional resources to logically sequence and achieve the broad job tasks.

As a training tool for the development of behavioral objectives, instruction, and examinations, the task inventory:

1. Serves as a central reference point for a learning analysis of the job tasks.

2. Shows the interrelationship of the various elements of instruction.

3. Forms a base for determining the instructional intent and the most efficient methods of insuring student achievement of the job tasks.

As a base line evaluation tool, the task inventory may be used to make the following determinations:

1. Does the instruction teach the training objectives which comprise the job tasks?

2. Do the tests examine the student's ability to perform behavioral objectives and job tasks?

3. Does the instructional system sufficiently and efficiently prepare the student for successful completion of his job after graduation?

It is expected that the major contribution of the Delphi Technique to training will be to reduce the time required to write and prioritize task lists. With the minimal time required to produce a task list using Delphi, it can easily be used to validate existing task lists.

This application of the Delphi Technique to job analysis resulted in a task inventory that met the predetermined criteria. The course selected for this experiment is representative of many military skill qualification courses. The applicability of the Delphi Technique to similar courses and courses of a more complex and theoretical nature bears further investigation.

SOURCES

1. Brown, B.; Cochran S. and Dalkey, N., *The Delphi Method, II: Structure of Experiments*; RAND Corporation. RM-5957, June 1969.

2. Coyle, H.F., Jr., *Project Total - To Teach All: An Inquiry into the Development of a Model for Identifying Unmet Needs in Urban Post Secondary Educational Offerings*. Akron University, Ohio, December 1973.

3. Cyphert & Gant; *The Delphi Tech: A Tool for Collecting Opinions in Teacher Education*. Journal of Teacher Ed., Fall 1970.

4. Dalkey, N.C.; *Predicting The Future*, RAND Corporation, October 1968.

5. Dalkey, Norman and Brown, Bernice; *Comparison of Group Judgment Techniques With Short-Range Predictions and Almanac Question*; RAND Corporation, May 1971.

6. Dalkey, Norman C.; *The Delphi Method: An Experimental Study of Group Opinion*; RAND Corporation, June 1969.

7. Dalkey, Norman C. and Rourke, Daniel L.; *Experimental Assessment of Delphi Procedures With Group Value Judgments;* RAND Corporation, February 1971.

8. Foley, John P., Jr., *Task Analysis for Job Performance Aids and Related Training*. Air Force Human Resources Lab., Brook AFB, Texas, 1973.

9. Grotelueschen, Arden D., *Influence of Cognitive and Affective Factors on Adult Learning: Three Experimental Studies*. University of Illinois, 1972.

10. Helmer, Olaf; *Social Technology*; RAND Corporation, February 1965.

11. Manis, Dane, Solomon, Warren., *Using Task Analysis to General Teacher Training Objectives and Activities and Evaluate Their Adequacy*. Central Midwestern Regional Educational Lab., St. Ann, Missouri, 1974.

12. *Procedural Manual for Task Analysis and Curriculum Guide Development*. Ohio State University, Columbus, School of Allied Medical Professions, Public Health Service (DHEW) Washington, D.C., 1972.

DEVELOPMENT AND VALIDATION OF CII PERFORMANCE CRITERION MEASURES AND COURSEWARE

W. G. Hoyt, A. K. Butler
System Development Corporation

and

D. J. Chesler
Navy Personnel Research and Development Center

INTRODUCTION

The purpose of the Computer Integrated Instruction (CII) project is to determine the feasibility of using computers on combatant ships for instruction and training management. The operational goal is to provide commanding officers of ships with additional or improved on-board computer capabilities to upgrade and maintain the combat readiness of their crews.

In December 1973, the System Development Corporation was tasked by the Navy Personnel Research and Development Center, San Diego, California to develop and validate CII in the shore-based environment prior to installation on a demonstration ship. The Navy is planning a shipboard test and evaluation of the CII system.

The minicomputer system used for this project is the NOVA 1200 equipped as shown in Figure 1. Aboard ship, the NOVA has been used for the Navy Maintenance and Material Management (3M) system, personnel administration, supply, and certain other ship-generated functions. The CII and training administration applications are in addition to and compatible with these uses of the minicomputer.

MINICOMPUTER HARDWARE/SOFTWARE SYSTEM

- NOVA 1200 CPU (32K WORDS CORE MEMORY)
- TWO FIXED HEAD DISK DRIVES (512K WORDS TOTAL CAPACITY)
- SIX MAGNETIC TAPE CASSETTE DRIVES
- LINE PRINTER (132 COLUMN)
- CARD READER (TABLE-MOUNTED)
- ASR 33 TELETYPE
- FOUR CRT TERMINALS

Figure 1. Minicomputer System Used for Computer Integrated Instruction (CII)

For this project, General Damage Control was identified as the subject area from which a subset of specific topics would be selected for courseware development. General Damage Control was selected as being representative of critical shipboard training requirements and the courseware developed for the selected areas was deemed to have broad applicability on-board ship. Further, attainment and maintenance of the proficiency in damage control is not achieved during normal day-to-day on-the-job operations. Damage control skills and knowledge are applied in emergency situations and the individual at the scene needs to know what to do and how to do it.

The document that served as the benchmark from which Damage Control topics would be selected was "Personnel Qualification Standard for Damage Control, Qualification Section 2, General Damage Control (PQS2A)." The areas of General Damage Control specified in PQS2A comprise the four theory sections, ten systems and one watchstation shown in Figure 2. From these, nine areas (*) were selected to provide a prototype instructional system for administration in the CII mode. The task was to develop approximately 30 hours of CII material from this General Damage Control domain, with 6 hours of this total allocated as tests for computer administration to shipboard personnel.

The CII material was developed as individualized instruction encompassing the multimedia modes of Programmed Instruction (PI), Audio Visual (AV) instruction, Audio (A) instruction, and Self-Study Guides (SSG). The procedure used to develop the CII curriculum, shown in Figure 3, followed established principles of course development. They evolved from an analysis as to what was to be developed through the presentation of completed instructional material to a representative target population, with appropriate review and revision cycles interspersed throughout the process. This presentation addresses each step in this procedure in the order in which it was performed, and closes with the conclusions and recommendations resulting from the entire effort.

SELECTION AND ANALYSIS OF SUBJECT MATTER AREAS

The initial task (Step 1 in Figure 3) was to select and analyze subject matter areas in General Damage Control. SDC started its analysis by examining the task statements and documentation in PQS2A. An example of these are the system and watchstation sections for the Oxygen Breathing Apparatus (OBA) system in Figures 4 and 5. In addition, discussions were held with personnel based in San Diego from the PQS Development Group, the Fleet Training Center Damage Control and Fire Fighting Schools, damage control personnel aboard ship, and the staff at NPRDC. The documentation analyzed is listed in the References.

CONTENTS

THEORY

* 2101	Damage Control
* 2102	NBC Defense
2103	First Aid and Rescue
* 2104	Safety Precautions

SYSTEMS

2201	Damage Control Organization
2202	Damage Control Communications
* 2203	Firemain
2204	Drainage
2205	Ventilation
* 2206	Fixed Damage Control Equipment
* 2207	Portable Damage Control Equipment
* 2208	Personnel Protective Equipment
2209	Mechanical Foam/Aqueous Film Forming Foam
* 2210	Oxygen Breathing Apparatus (OBA)

WATCHSTATIONS

* 2301	Shipboard Damage Control

Figure 2. Theory, Systems and Watchstation Areas of the Personnel Qualification Standard for Damage Control, Qualification Section 2, General Damage Control, NAVTRA 43119-2A (Asterisk denotes selected areas)

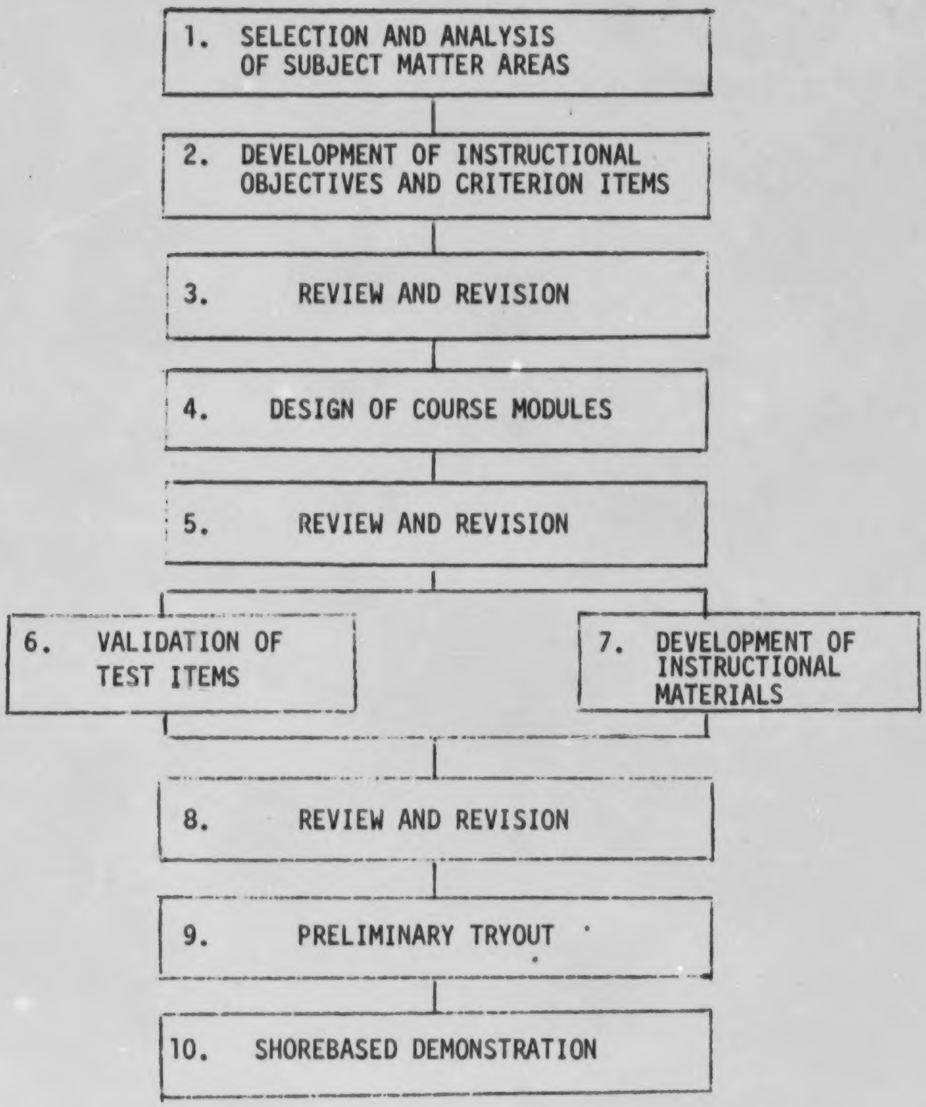

1. SELECTION AND ANALYSIS OF SUBJECT MATTER AREAS
2. DEVELOPMENT OF INSTRUCTIONAL OBJECTIVES AND CRITERION ITEMS
3. REVIEW AND REVISION
4. DESIGN OF COURSE MODULES
5. REVIEW AND REVISION
6. VALIDATION OF TEST ITEMS
7. DEVELOPMENT OF INSTRUCTIONAL MATERIALS
8. REVIEW AND REVISION
9. PRELIMINARY TRYOUT
10. SHOREBASED DEMONSTRATION

Figure 3. CII Course Development Steps

2210 OXYGEN BREATHING APPARATUS (OBA) SYSTEM

2210.1 Explain the function(s) of the OXYGEN BREATHING APPARATUS (OBA) SYSTEM as stated in Hull Maintenance Technician 3&2 (NAVEDTRA 10573) and Basic Military Requirements (NAVEDTRA 10054).

.11 Refer to a standard print of this system during the rest of this discussion.

2210.2 SYSTEM COMPONENTS - GENERAL

Discuss the designated items for each component listed below:

A. Explain the function(s) of the component in terms of what it does for the system.
B. Show or describe the actual physical location of this component.
C. Describe the sources of power.
D. Describe the color designations.

		A	B	C	D
.21	Adjustable head harness	X	X		
.22	Face piece	X	X		
.23	Vent valve	X	X		
.24	Breathing tubes	X	X		
.25	Central casting	X	X		
.26	Timer	X	X	X	
.27	Canister stop	X	X	X	
.28	Breathing bag	X	X		
.29	Bail	X	X		
.210	Hand screw	X	X		
.211	Adjustable shoulder harness	X	X		
.212	Shoulder harness "D" rings	X	X		
.213	Adjustable waist strap	X	X		
.214	Waist strap ring	X	X		
.215	Canister	X	X	X	X

2210.3 COMPONENT PARTS

Discuss the designated items for each component part listed below:

A. Explain the function(s) of the component part in terms of what it does for the component.
B. Describe the functional location of this component part within the component.
C. Describe the physical location of this component part within the component.
D. Describe the physical location of this component part within the circuit.

Figure 4. PQS2A, 2210 Oxygen Breathing Apparatus (OBA)

2301 WATCHSTATION - SHIPBOARD DAMAGE CONTROL

2301.1 OPERATING INSTRUCTIONS

For the operating instructions listed below.

A. Describe the sequence of steps of this procedure.
B. Explain the reasons for each step of this procedure.
C. Discuss the control/coordination required while performing this procedure.
D. Discuss the communications that must be established and/or utilized.
E. Discuss the parameter indication(s) that must be monitored.
F. Discuss the safety precautions that must be observed.
G. Perform the steps of this procedure.
H. Perform the steps of this procedure when practicable.
I. Perform the steps of this procedure using the compartment checkoff list.

* * * *

		A	B	C	D	E	F	G	H	I
.16	Demonstrating proper securing of ports, hatches, doors, scuttles and deck drains	X	X	X				X		
.17	Obtaining, checking, donning and stowing an OBA and canister	X	X			X	X	X		
.18	Demonstrating OBA tending signals	X	X			X	X	X		
.19	Demonstrating self-protection procedures used in NBC defense	X	X			X		X		
.110	Reading pocket dosimeter	X	X			X		X		
.111	Obtaining, checking, donning and stowing a Mark V protective mask	X	X			X		X		

* * * *

2301.2 NORMAL OPERATIONS

For the conditions or evolutions listed below:

A. Define the parameters monitored.
B. Explain how the parameters change.
C. Proper stowage.

		A	B	C
.21	While setting material conditions of readiness	X	X	
.22	While using an Oxygen Breathing Apparatus (OBA)	X	X	X
.23	While using a Mark V protective mask	X	X	X

* * * *

Figure 5. PQS2A, 2301 Watchstation

PQS Number	Subject Area
2101.2	Compartmentation and Watertight Integrity
2101.3	Fire and Fire Fighting
2101.4	Battle Damage Types
2101.5	Battle Damage Repair
2102.1	Nuclear Defense
2102.2	Biological Defense
2102.3	Chemical Damage
2104.1	Safety Precaution - General
2203.25	Magazine Sprinking System
2203.27	Water Washdown System
2206.21	Fixed CO2 System
2206.26	Twin Agent System (AFFF and PKP)
2207.21	Firehose
2207.22	All-Purpose Nozzle
2207.23	Low-Velocity Fog Applicator
2207.24	Portable Water Motor Proportioner (FP-180)
2207.25	Navy Pickup Unit (NPU) Nozzle
2207.26	Foam Nozzle
2207.27	CO2 Fire Extinguisher
2207.28	Dry Chemical (PKP/ABC) Extinguisher
2207.29	P-250 Internal Combustion Engine Pump
2207.210	Electric Submersible Pump
2207.211	Eductors (Peri or Single Jet)
2207.212	Portable Electric Blower
2207.213	Portable Pneumatic Blower
2207.214	Flood and Battle Lanterns
2207.215	Eductor Discharge Hose
2208.21	Proximity Suit and Gloves
2208.22	Impregnated (Permeable) Protective Clothing
2208.24	Mark V Protective Mask
2208.28	CO2 Inflatable Lifejacket
2208.29	Casualty Dosimeter (DT-60/PD)
2208.210	Pocket Dosimeter (IM-132/PD)
2210	Oxygen Breathing Apparatus

Figure 6. PQS2A Topic Selection

TAIS No. 1006

MODULE CII OBA
UNIT OBA

TRAINING ANALYSIS INFORMATION SHEET

1. TASK IDENTIFICATION: 6.0

2. TASK: State the proper sequence for putting the OBA in operation (ready position) from the standby (patrol) position.

3. CONDITIONS: Given a list of procedural steps on putting the OBA in operation from the standby position, select the order in which they are performed.

   Given multiple-choice and constructed response questions, provide correct response.

4. STANDARD: No errors.
5. TASK ANALYSIS:

TASK ELEMENTS	PREREQUISITE KNOWLEDGE OR SKILL REQUIREMENTS	SUPPLEMENTAL TRAINING MATERIAL	REFERENCES
6.1 Identify the procedures for putting the OBA into operation from a standby position: a. Push in canister stop and hold. b. Tighten handscrew until tight, then let go canister stop. c. Fire chlorate candle. d. Don facepiece. e. Breathe normally. f. Set timer. g. Check and adjust OBA, if necessary to clear fogging lenses.	6.1 Know component parts of OBA.	Pictures or drawings of the OBA to show sequential steps of putting the OBA in operation.	1. NAVPERS 10054-C Chapter 11, Pages 165 through 169. 2. NAVPERS 10571-F, Chapter 7, Pages 92 through 95. 3. NAVTRA 9880-1017 through 1031; 9880-1034; 9880-1068 through 1075. 4. NAVSHIPS 393-0573, Page 28; Pages 32 through 39; Pages 72 through 74.

Figure 7. Training Analysis Information Sheet (TAIS) - OBA

TAIS No. 1006

MODULE CII OBA
UNIT OBA

CRITERION AND ENABLING OBJECTIVES

TASK IDENTIFICATION: 6.0

TASK ELEMENTS: 6.1

CRITERION OBJECTIVE(S)	ENABLING OBJECTIVE(S)
6.1 Given a list of procedures for putting the OBA in operation (ready position) from the standby (patrol) position, but with the procedures in a scrambled order, the student can state the correct order in which these procedures are performed. The correct order is: a. Release (push in) canister stop and hold. b. Tighten handscrew until canister is sealed (tight) and let go canister stop. c. Fire chlorate candle. (1) Turn swivel 180° and push down. (2) Pull lanyard out (away) from you. (3) Check that candle has fired. (a) Cotter pin on lanyard. (b) Breathing bag fills and presence of smoke. d. Don facepiece. (1) Check air tightness. e. Breathe normally. (1) Release excess pressure in breathing bag. (a) Press starter (vent valve). (b) Squeeze breathing bag gently with elbows until pressure is released (bag is three-fourths filled). f. Set timer. (1) 60 minute normal supply of oxygen. (2) 45 minute use, 15 minute safety factor.	6.1.1 Select from a list that the first step in going from the standby to the operation (ready) position on the OBA is: PUSH IN CANISTER STOP AND HOLD. 6.1.2 Pick from a list the result if the canister stop is not pushed in: THE CANISTER IS PREVENTED FROM BEING SEALED AGAINST THE MAIN VALVE ASSEMBLY. 6.1.3 Select from a list two indications that the chlorate candle has not fired. The two are: 1. Cotter pin is missing from lanyard. 2. Breathing bag does not fill. 6.1.4 Select from a list the action to take to release excess pressure in the breathing bag: PRESS STARTER VALVE AND SQUEEZE BREATHING BAG. 6.1.5 State 60 MINUTES as being the time that the oxygen supply in a canister normally lasts. 6.1.6 State 45 MINUTES as being the time duration that the timer is normally set. 6.1.7 Select from a list the steps to take if fogging on the lenses of the OBA facepiece occurs: CHECK BREATHING BAG. 6.1.8 Select from a multiple-choice list the action to take if you have been in a smoke-filled compartment for 30 minutes and the lenses in the facepiece of the OBA start fogging as you inhale: MOVE OUT TO UNCONTAMINATED AIR.

Figure 8. Criterion and Enabling Objectives - OBA

Selection was based upon skills and knowledge that are truly representative of General Damage Control, that can be converted into learning objectives, and that contain performance-oriented tasks amenable to individualized instruction. The specific topics in these areas for which instruction was developed are shown in Figure 6.

For each subject area selected, Training Analysis Information Sheets (TAIS) were prepared. TAISs contain the task statement, conditions and standards, and the results of and sources for the task analysis. A representative TAIS for the OBA is shown in Figure 7.

DEVELOPMENT OF INSTRUCTIONAL OBJECTIVES AND CRITERION ITEMS

Development of instructional objectives and criterion items was the next task (Step 2, Figure 3) in CII courseware development. Instructional objectives serve as a base from which instructional material is developed and lead directly to the development of criterion test items. Criterion objectives were developed for each task element specified in the TAIS. The criterion and enabling objectives for TAIS No. 1006 for the OBA is shown in Figure 8. Note that the TAIS number and task element number are the same here as for the TAIS in Figure 7. This provides an audit trail between the TAIS, criterion and enabling objectives, and also the test items.

Criterion and enabling test items were keyed directly to the criterion and enabling instructional objectives. Guidelines used were:

- Test items are performance oriented and require the student to demonstrate skills and knowledge directly related to the criterion objectives.
- Each test item elicits measurable behavior.
- The structure of the test item is positively oriented.
- When multiple-choice items are used, there are at least four alternatives.
- The test is amenable to on-computer (CRT) presentation, or CRT presentation plus a simple off-line exhibit.

Figure 9 shows a test item worksheet. The correct answers are asterisked in parenthesis, or underlined, depending on whether they are multiple-choice, constructed response, or matching test items.

TAIS No. 1006

MODULE CII OBA
UNIT OBA

TEST ITEMS

TASK IDENTIFICATION: 6.0

TASK ELEMENTS: 6.1

CRITERION ITEM(S)	ENABLING ITEM(S)
6.1 A. You have the OBA on in the standby (patrol position) and you are going to put the OBA in operation (ready position). The first step, Step 1, is which one of the following? (Select a letter) a. Put on facepiece and check air tight. b. Set timer. c. Tighten handwheel until canister is seated (tight) and let go canister stop. d. Fire chlorate candle. e. Release (push in) canister stop and hold. (e) B. You have the OBA on in the standby (patrol position). Step 3 of the procedure is: Fire chlorate candle. Which one of the following is the next step, Step 4? a. Put on facepiece and check air tight. b. Breathe normally and relieve excess pressure in breathing bag. c. Tighten handwheel until canister is seated (tight) and let go canister stop. d. Release (push in) canister stop and hold. (a)	6.1.1 The first step in going from the standby position to the operation (ready) position on the OBA is which of the following? *a. Release (push in) canister stop and hold. b. Put on facepiece. c. Fire chlorate candle. d. Manual start. 6.1.2 In going from the standby to the operation (ready) position on the OBA, if the canister stop is not pushed in and held, it will have which one of the following effects: a. Lock the timer so it cannot be set. *b. Prevent canister from being seated against the main valve assembly. c. Puncture the canister and make it inoperable. d. Require a manual start of the OBA.

Figure 9. Test Item Worksheet - OBA

The TAIS, objectives, and test items were reviewed and approved (Step 3 in Figure 3) by representatives of the Fleet Training Center Damage Control and Fire Fighting Schools, San Diego, NPRDC, COMNAVSURFLANT, ship personnel, and Chief of Naval Operations.

DESIGN OF COURSE MODULES

Courseware was organized as shown in Figure 10. Each of the course modules contained a pre-test and post-test taken on-computer and a number of lessons taken off-line in the form of Programmed Instruction (PI), Audio Visual (AV) instruction, Audio (A) instruction (with the equipment), or as Self-Study Guides (SSG). PI lessons contained the five elements shown at the bottom of Figure 10. Element 1 contained the lesson objectives and element 2, the lesson content. The review element contained the list of topics within the lesson and their location. The criterion test element included a scoring key for self-scoring the test. The self-evaluation element enabled the student to evaluate his test results so as to determine his preparedness for the on-computer module test.

VALIDATION OF TEST ITEMS

The validation of test items (Step 6, Figure 3) was a two-part process. The module tests, containing 386 test items, were administered at the Fleet Training Center, San Diego, to 28 subject matter experts and instructors from the Fire Fighting School (27 subjects) and the Damage Control School (1 subject), and to 30 novice subjects, who were newly arrived from boot camp and awaiting further assignment. The tests were administered off line in hard copy format as they would be displayed on the computer CRT. A sample test item sheet is shown in Figure 11. Note that the audit trail is maintained by the header on the test item for both the TAIS and PQS2A, e.g., 29, 1006.1A, 2210 means test item number 29, from TAIS 1006, for the PQS2A 2210 module. The accompanying off-line exhibit for the test items is shown in Figure 12.

After the first tryout, test items were modified primarily to change the presentation format of test items dealing with procedures from a scrambled order to the direct, multiple-choice format shown in Figure 11. The revised module tests, consisting of 405 test items, were administered to 10 "expert" subjects and 10 novice subjects awaiting assignment from the Fleet Training Center. The 10 "expert" subjects were five Damage Control School instructors, three welder instructors and two non-instructors--one a Hull Technician and the second, an Operations Specialist. Only the five Damage Control School instructors could be considered as subject matter experts in damage control, as the remaining five were instructors or working in areas other than damage control. This was especially true for the Operations Specialist, whose test scores were within the range of novice subjects.

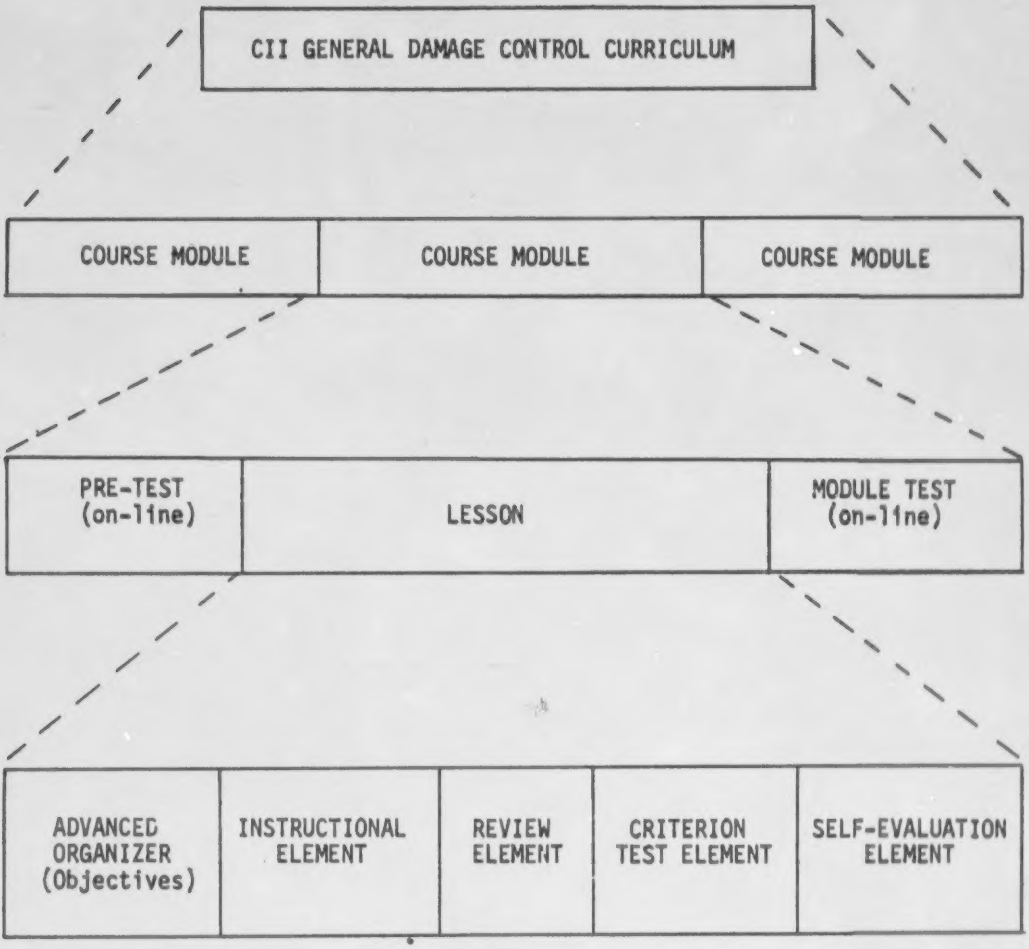

CII GENERAL DAMAGE CONTROL CURRICULUM
COURSE MODULE
COURSE MODULE
COURSE MODULE
PRE-TEST (on-line)
LESSON
MODULE TEST (on-line)
ADVANCED ORGANIZER (Objectives)
INSTRUCTIONAL ELEMENT
REVIEW ELEMENT
CRITERION TEST ELEMENT
SELF-EVALUATION ELEMENT

Figure 10. CII Courseware Organization

29 1006.1A 2210

YOU HAVE THE OBA ON IN THE STANDBY (PATROL POSITION) AND YOU ARE GOING TO PUT THE OBA IN OPERATION (READY POSITION). THE FIRST STEP, STEP 1, IS WHICH ONE OF THE FOLLOWING? (SELECT A LETTER)

A. PUT ON FACEPIECE AND CHECK AIR TIGHT
B. SET TIMER
C. TIGHTEN HANDSCREW UNTIL CANISTER IS SEATED (TIGHT) AND LET GO CANISTER STOP
D. FIRE CHLORATE CANDLE
E. RELEASE (PUSH IN) CANISTER STOP AND HOLD

YOUR ANSWER: __________

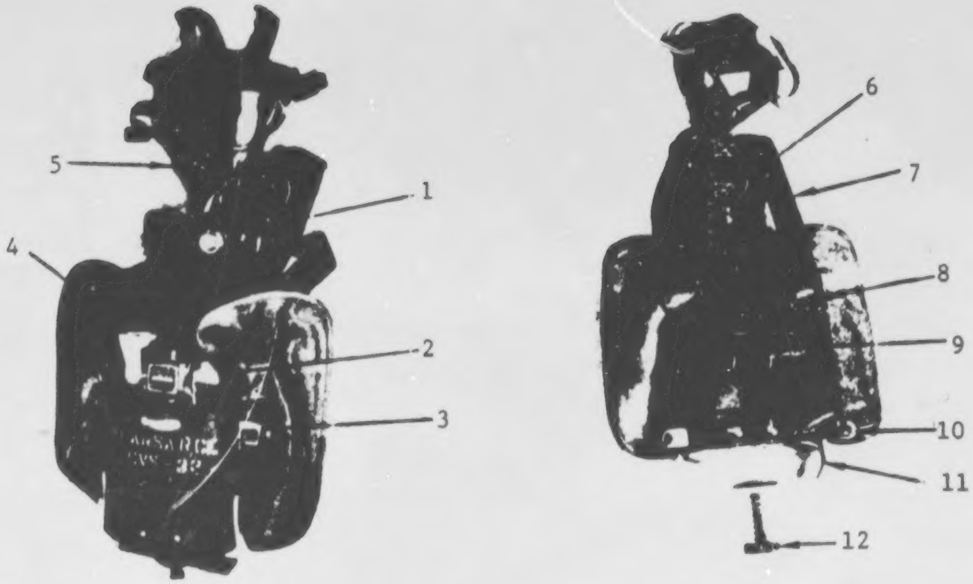
30 1006.1B 2210

YOU HAVE THE OBA ON IN THE STANDBY (PATROL POSITION) AND YOU ARE GOING TO PUT THE OBA IN OPERATION (READY POSITION). STEP 3 OF THE PROCEDURE IS: FIRE CHLORATE CANDLE. WHICH ONE OF THE FOLLOWING IS THE NEXT STEP, STEP 4? (SELECT A LETTER)

A. PUT ON FACEPIECE AND CHECK AIR TIGHT
B. BREATHE NORMALLY AND RELIEVE EXCESS PRESSURE IN BREATHING BAG
C. TIGHTEN HANDSCREW UNTIL CANISTER IS SEATED (TIGHT) AND LET GO CANISTER STOP
D. RELEASE (PUSH IN) CANISTER STOP AND HOLD

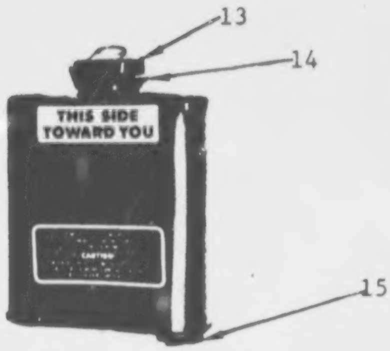
YOUR ANSWER: __________

Figure 11. Test Item Presentation - OBA


5
1
4
2
3
6
7
8
9
10
11
12

OBA - Front View

OBA - Rear View


13
14
THIS SIDE TOWARD YOU
CAUTION
15

OBA Canister

Figure 12. Off Line Exhibit - Module 2210 OBA

Figure 13 displays the results from the two test administrations, using only the scores of the five Damage Control instructors as subject matter experts in the second administration. The reversal in modules 2206 and 2210 are not unexpected, as 2206 and 2210 are taught (including practical application) at the Fire Fighting School and only demonstrated at the Damage Control School. The results showed that the module tests clearly differentiated between novices and subject matter experts. Further, the content validity of the test items had been substantiated, first by individual subject matter experts who reviewed the test items, and second, by the groups of subject matter experts, who, on both the first and second test administrations, were asked to mark problem areas as they took the tests, and were later asked to define exactly what the problems were.

DEVELOPMENT OF INSTRUCTIONAL MATERIALS

In the development of CII materials (Step 7 in Figure 3), four types of instruction were specified for the CII curriculum: Programmed Instruction (PI), Audio Visual (A/V) instruction, Audio (A) instruction used in conjunction with the equipment, and Self-Study Guides (SSG).

Programmed Instruction (PI) contains instructional material in a planned sequence that provides immediate feedback to the student, permits him to proceed at his own pace, and requires maximum student participation.

A task, as specified in the TAIS, became the basic instructional unit. A series of frames were prepared in conjunction with each enabling or criterion test item in the lesson sequence. Each frame was designed to perform one or more of the following functions:

- Present procedures, content information, examples, test items, or instructions to the student.
- Evaluate a student response as correct or incorrect.
- Provide appropriate feedback.
- Decide on the next action to be taken, i.e., proceed in sequence or branch to another point in the lesson.

In developing these frames, the basic ground rules and principles of instructional programming were adhered to, i.e.,:

- Let the student know where he is going and why that is important.
- Inform the student of his performance over sets of subgoals.

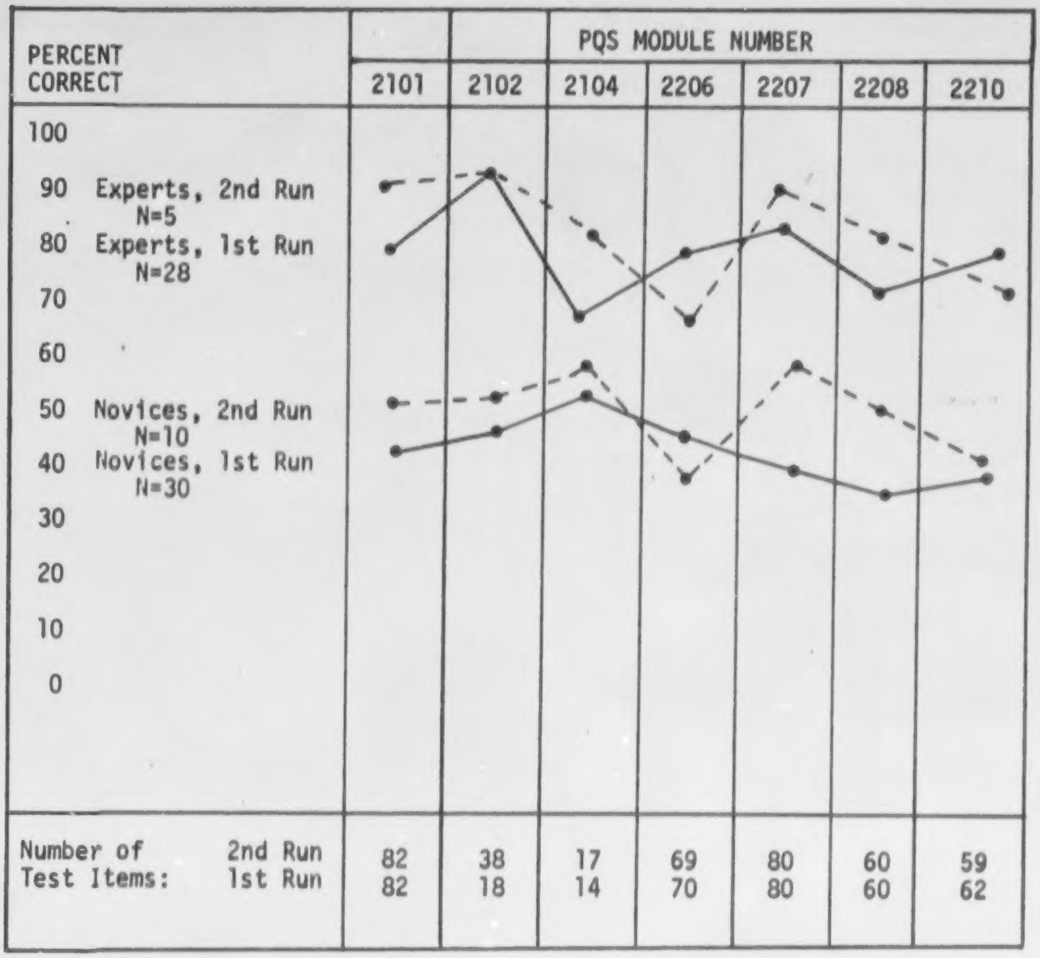


Figure 13. Results of 1st and 2nd Test Administrations

- Provide clear instructions--avoid ambiguity of what is required.

- Keep information and feedback as straightforward and concrete as possible.

This process was repeated until the frames for all the tasks in a specific lesson were completed. These completed frames were incorporated into a Programmed Instruction booklet. The organization of each PI booklet was the same, as was shown in Figure 10.

Audio-Visual (AV) instruction embodies 35mm slides with synchronized narrative on a cassette tape. For each A/V lesson, slides were prepared and a script written, based upon the TAIS, criterion and enabling objectives, and test items. The equipment used to prepare the tapes and display the lesson was the Wollensak 2548 AV recorder/player. A/V differed from PI in these ways:

- Greater emphasis was placed on showing the steps to procedures rather than on discussing them.

- Procedures were broken down into detailed visual sequences.

- When questions were asked, students could not examine an answer printed in the following frame; instead, a 5- to 20-second pause allowed for answer review.

Audio (A) instruction is a taped narration used in conjunction with specific equipment. This instructional mode was used with the Oxygen Breathing Apparatus, Mark V protective mask, and the Twin Agent System. A script was prepared for each lesson based upon the TAIS, criterion and enabling objectives and test items. Each script took the student through the procedures and checks necessary to operate the equipment; a prerequisite was that he had taken the PI and/or A/V lessons on the equipment. This mode of instruction was designed to:

- Provide hands-on experience with the equipment.

- Provide immediate feedback following a step or procedure.

- Permit timed practice sessions where such were deemed important.

The Self-Study Guides (SSG) developed contain the tasks (objectives) from PQS2A, the pages from Navy publications referenced in PQS2A, test items and scoring key, and a method of evaluating performance. The two Self-Study Guides prepared were 2102, NBC Defense, and 2104, Safety Precautions. The Self-Study Guide, without the test items and scoring and method of evaluating performance, is the method normally used by the individual in preparing for the Qualification Standard (PQS).

The total CII course structure is shown in Table 1. The modules are designed as independent units which can be taken in any sequence desired. Lessons are PI (Programmed Instruction), A/V (Audio-Visual), A (Audio), or SSG (Self-Study Guide). Where all three media are used, e.g., 2210 Oxygen Breathing Apparatus, PI is used for basic thorough learning in detail of all the material, followed by the A/V lesson, which visually reviews and reinforces primarily operational procedures. An Audio lesson in which the individual actually puts on the OBA, including firing the OBA canister and making prescribed checks as he goes along, completes the instructional sequence.

REVIEW AND REVISION — CII COURSEWARE

Navy subject matter experts reviewed the CII courseware (Step 8 in Figure 3). Differences between Navy publications and actual practice aboard ship were identified and modifications made to the CII courseware accordingly. At this time a revised PQS2, PQS2A, was published which changed the organization, format and content of the CII materials. These changes were also incorporated.

PRELIMINARY TRYOUT

The purpose of the preliminary tryout (Step 9, Figure 3) was to determine whether the instructional materials are clear, whether the sequence of the lesson is appropriate, whether additions or revisions are required, and student reaction to CII.

Testing for each module was done off line using hard copy printouts from punched cards of the test items as they would appear on the CRT. From the pool of validated test items, two test versions (A and B) for each module were prepared. Each version contained 298 test items. The same items that appeared in both versions were altered by varying the order of alternatives presented or by selecting different steps in a procedure or a different component part. For certain items where there was only one answer or only one job-oriented way of presenting the item, e.g., identify the canister stop on the OBA, the same item was included, as is, in both versions. Off-line course materials consisting of pictures or diagrams of the equipment were used to put the individual into the job situation when the test item was presented on the CRT.

The four novice subjects who participated in the preliminary tryout had recently completed boot camp and were currently assigned to the Fleet Training Center, San Diego. Each individual took the pre-test (Version A), the instructional material, and post-test (Version B) for each module. Pre-test scores averaged 55%, post-test scores 85%. Reaction was highly favorable. All four liked CII very much and thought this method of instruction very effective. They liked the individual study: "You work at your own rate so you learn more." "You had a better picture of

Table 1. CII General Damage Control Course Structure

Lesson	PQS No.	Module/Topic	Type	TAIS Nos.
		PQS 2101 DAMAGE CONTROL:		
1	2101.2	Compartmentation and Watertight Integrity	PI	3001-3008
2	2101.3	Fire and Firefighting	PI	3101-3105
3	2101.4	Battle Damage Types	PI	3201-3205
4	2101.5	Battle Damage Repair	PI	3301-3305
		PQS 2102 NBC DEFENSE	SSG	
		PQS 2104 SAFETY PRECAUTIONS	SSG	
		PQS 2203 FIREMAIN SYSTEM:		
1	2203.27	Water Washdown/Magazine Sprinkler Systems	PI	5201-5204
		PQS 2206 FIXED DAMAGE CONTROL EQUIPMENT:		
1	2206-21	Fixed CO2 System	PI	5101-5105
2	2206.26	Twin Agent System (AFFF and PKP)	PI	5301-5307
3		Twin Agent System (AFFF and PKP)	A/V	5401-5406
4		Twin Agent System (AFFF and PKP)	A	5501
		PQS 2207 PORTABLE DAMAGE CONTROL EQUIPMENT:		
1	2207.2	Hoses, Nozzles and Foam Equipment	A/V	6101-6108
2	2207.2	Extinguishers	PI	6001-6010
3	2207.2	Pumps/Eductors	A/V	6201-6210
4	2207.2	Blowers and Lanterns	PI	6301-6302
		PQS 2208 PERSONNEL PROTECTIVE EQUIPMENT:		
1	2208.2	Protective Clothing	A/V	2501-2504
2	2208.24	Mark V Protective Mask (ND MK-V)	PI	2001-2010
3		Mark V Protective Mask (ND MK-V)	A	2101
4	2208.28	CO2 Inflatable Lifejacket	A/V	2601-2604
5	2208.29	Casualty Dosimeter (DT-60/PD)	PI	2301-2305
6	2208.210	Pocket Dosimeter (IM-143/PD)	PI	2201-2206
		PQS 2210 OXYGEN BREATHING APPARATUS (OBA):		
1	2210	Oxygen Breathing Apparatus (OBA) Type A-3	PI	1001-1009
2		Oxygen Breathing Apparatus (OBA) Type A-3	A/V	1101-1106
3		Oxygen Breathing Apparatus (OBA) Type A-3	A	1201

how to do it and why you have to do it." "I have learned a great amount of which I think most of it will stay with me throughout my career."

Based upon the preliminary tryout, revisions were made to clarify course content and presentation.

SHORE-BASED DEMONSTRATION

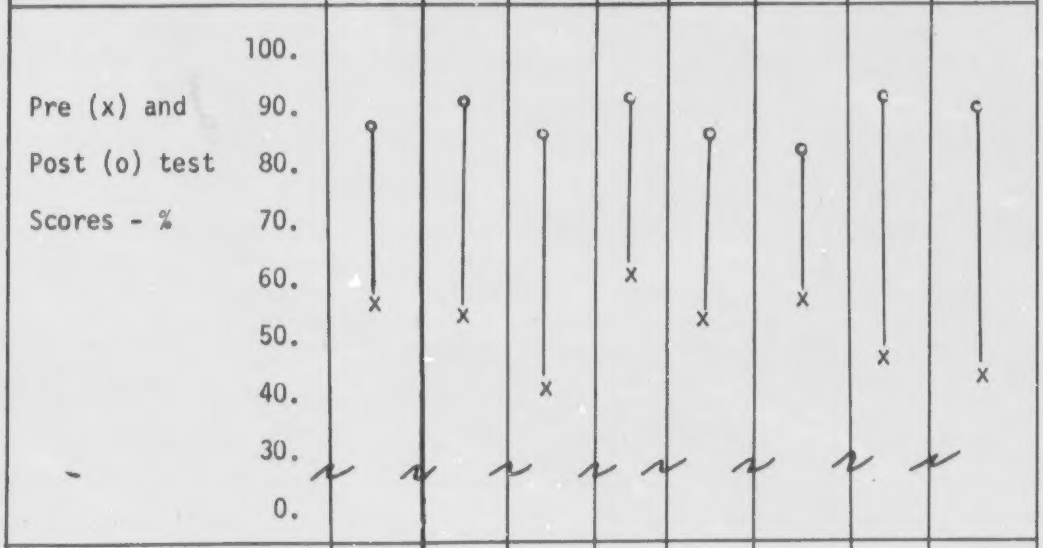
The purpose of the shore-based demonstration (Step 10 in Figure 3), replicating shipboard use, was to demonstrate the effectiveness of the CII training program. This included the CII course materials and procedures and the CII computer program for on-line testing. Additionally, the demonstration also involved the testing of the integration of the instructional materials with on-line computer testing and CRT presentation of prescriptives. The shore-based demonstration, conducted at the SDC facilities in Falls Church, Virginia, involved 12 Navy subjects--8 seamen recruits and 4 firemen recruits--who had recently completed boot camp and were moving to their first assignment.

The recruits were given orientation as to the purpose of the shore-based demonstration. After filling out a background questionnaire, they were given "Introduction to CII Instruction" for a general familiarization of the total CII program. This included data on PQS2, the structure of the CII curriculum, and procedures for taking each method of instruction (PI, A/V, A and SSG) and the on-line module tests. Students were also given a handout titled "Index to Lessons and Self-Study Guides" by which they could monitor their progress in taking the CII tests and associated materials.

Each subject started with a module pre-test. If he passed, he was credited with the PQS2 points for the module and went on to the next module pre-test. If he failed the module pre-test, he took the CII lessons for the module, including the review and self-evaluation elements shown in Figure 10. When he was ready, he took the post-test on line. If he failed, he retook the appropriate lessons and recycled until he passed the post-test. This process was repeated for each module.

The passing score was 80% for each module plus passing of every critical test item. Critical test items, e.g., the canister stop on the OBA, were those that would result in loss of life or serious damage if not known.

Results for the pre- and post-tests are shown in Figure 14. The subjects who passed the pre-test, shown on the next to bottom line, are not included. Pre-test averages ranged from 42% to 63%, post-test scores from 82% to 90%. Five subjects required recycling (taking the module again before they passed the post test) on module 2101, five on module 2207, and one each on module 2206 and 2208.



(No. of Test Items)	PQS 2A Module Tests 2101 (52)	2102 (29)	2104 (13)	2203 (6)	2206 (43)	2207 (62)	2208 (43)	2210 (44)
No. completing Module	11	10	7	5	9	8	10	4
No. passing pre-test	0	2	4	7	0	0	0	0
No. recycling	5	0	0	0	1	5	1	0

Figure 14. Module Test Scores - Shore Based Demonstration

Not all subjects attempted every CII module. The module that showed the greatest gain was 2210, the OBA, from 43% to 89%. None of the four subjects who went through the OBA required recycling, even though three of the four required recycling on other modules. The OBA module included all three types of instruction: Programmed Instruction (P/I), Audio-Visual (A/V), and Audio (A) with the equipment.

At the end of the shore-based demonstration, subjects were asked to fill out the CII Debriefing Questions. Responses to the questions showed that the attitude of the subjects toward CII was extremely favorable. They liked the course and self-paced structure of the materials, and thought this was a very effective method of learning about General Damage Control. Typical comments are shown in Figure 15.

CONCLUSIONS AND RECOMMENDATIONS

The conclusions and recommendations which follow stem from the results obtained and problem areas identified and resolved successfully within and during the conduct of the CII project. The application of standards, proficiency testing, development of course materials, and new equipment training are not uniquely CII problems but apply throughout the Navy and other services as well.

STANDARDS FOR GENERAL DAMAGE CONTROL

The Personnel Qualification Standard for Damage Control, Qualification Section 2, General Damage Control, NAVTRA 43119-2A provides a set of generalized tasks rather than standards for the General Damage Control area. The development of standards is left to each Commanding Officer, each division head, and each supervising officer or petty officer. Their sets of standards may not only differ, but become more specific down through the chain of command. Considering the numbers of ships and personnel involved, the development of unique sets of standards is a time-consuming and difficult task. Further, with so many different standards, it is a major problem to accurately assess an individual's qualifications.

SDC developed a set of standards tied to specific learning objectives for the General Damage Control modules specified in the CII course for PQS2A. Damage Control subject matter experts reviewed these standards and considered them valid, as did instructors at the Fire Fighting and Damage Control Schools, who have extensive fleet experience. These standards also were reviewed and considered valid by COMNAVSURFLANT and ship personnel who are close to the current operating situation.

SDC recommends that the standards for General Damage Control developed for the CII course be considered for dissemination and application throughout the Navy to: (1) lessen the workload of fleet personnel;

You can't go to another lesson without learning the one you are studying.

You have time to learn the material.

It was fun, easy to learn from.

This course was very interesting and taught me many things which could save my life and others.

It was enjoyable cause no one made you go fast and it was for my benefit.

I enjoyed the program and wish there was more and also it was the best way to learn damage control.

I liked it very much. A great way to find out what information I did not know and learn it quickly.

Effective. I learn very much. Better than a teacher.

I think it is a good course and I think everybody going on a ship should take it.

I think the best part was taking the computerized tests because it gave me a feeling of accomplishment when I passed.

If you failed, you got another chance at it.

I learned more here in two weeks than I did in nine weeks at Great Lakes, they should have this system in all of the boot camps.

At first, I disliked the idea, but later I found out that all Navy personnel should know what we've learned.

Figure 15. Comments of Navy Students Taking the CII Course

(2) provide guidelines for testing personnel; and (3) provide a basis for using one set of standards throughout the Navy for equipments covered by the CII course.

PROFICIENCY LEVEL TESTING

Navy personnel aboard ship have varied backgrounds in shore-based training, shipboard training, practical experience, and instructor duties in the area of General Damage Control. Because only a limited amount of instructor and student time is available for training, the training needs of ship personnel should be positively identified, so that instructor and student time can be directed toward those areas in which proficiency is below desired levels.

SDC recommends that the test modules developed for the CII course be considered for use throughout the Navy to assist supervisory personnel in identifying personnel who are below or above the desired level of proficiency. Training can then be scheduled for personnel with low proficiency, while personnel whose proficiency is above the required level (and the instructors) can pursue other training requirements.

CII GENERAL DAMAGE CONTROL COURSE MATERIALS

The CII course, which covers a wide range of General Damage Control equipment, is effective in teaching Navy personnel the procedures, knowledge, skills, safety requirements, and other information required to satisfactorily operate and maintain the equipment covered. Damage Control subject matter experts have reviewed the CII course materials and consider them valid. Navy personnel respond favorably to the CII instruction methods, particularly the five-step approach used to attain proficiency: (1) Pre-test to determine areas of deficiency; (2) Programmed Instruction (PI) to provide the skills and basic knowledge in required procedures; (3) Audio-Visual (A/V) instruction to reinforce the learning and display step-by-step procedures required in a specific situation; (4) Audio (A) instruction to provide practical, hands-on experience with equipment; and (5) Post-test to demonstrate attainment of proficiency.

SDC recommends that: (1) CII course materials be disseminated Navywide to assist Navy supervisory personnel in providing training in the area of General Damage Control; and (2) the methodology and methods of instruction developed in the CII course be considered for application to similar areas where day-to-day operations of the ship do not normally exercise the procedures, skills, and knowledge required and where maintenance is a problem.

NEW EQUIPMENT TRAINING

Because new equipment, such as the Twin Agent System, is continually being introduced to the Navy, a training program is necessary to accompany or precede the operational introduction of these equipments. Equipment manufacturers' publications are not sufficiently detailed, organized, or situationally oriented to adequately train personnel to attain or maintain proficiency. Requiring every supervisor and instructor to assimilate, interpret, develop and validate the same operational and training program requires duplicate man-hours, obviates cost-effectiveness, and fails to resolve the standardization problem. Such a training program would immediately benefit Navy supervisors and instructors who are responsible for operating and maintaining the equipment and for training other Navy personnel. In addition, trained personnel can be on-board when the equipment is installed and checked out to ensure that the equipment can be operated and that it can perform its specified functions.

SDC recommends that the methodology, principles, and applications used for CII course materials be used to develop training courses and standards on new equipment to provide familiarization and proficiency training to supervisory personnel aboard ship and to school instructors who will be responsible for acceptance, operation, and training on new equipment. These training courses, in turn, would be easily adaptable to either shipboard or shore-based training, on an individual or group basis, for all Navy personnel requiring such training.

REFERENCES

1. Personnel Qualification Standard for Damage Control, Qualification Section 2, General Damage Control, NAVTRA 43119-2A, Chief of Naval Training, 1974.

2. Handbook on Personnel Qualification Standards, Implementation Procedures, NAVTRA 43100-1, Chief of Naval Training, 1973.

3. Damage Control: Practical Damage Control, NAVSHIPS Technical Manual, Chapter 9880, Section II, Naval Ship Systems Command, August 1973.

4. Shipboard BW/CW Defense and Countermeasures, NAVSHIPS Technical Manual, Chapter 9770, Naval Ship Systems Command, 1968.

5. Damage Control – Stability and Buoyancy, NAVSHIPS Technical Manual, Chapter 9881, Naval Ship Systems Command, 1967.

6. Fire Fighting – Ship, NAVSHIPS Technical Manual, Chapter 9930, Naval Ship Systems Command, 1967.

7. Basic Military Requirements, NAVPERS 10054-C, Bureau of Naval Personnel, Third Revision, 1970.

8. Hull Maintenance Tech 3 and 2, Rate Training Course, NAVTRA 10573, Naval Training Command, 1972.

9. Damage Controlman 1 and C, Rate Training Manual, NAVPERS 10572-D, Bureau of Naval Personnel, Revised 1965.

10. Damage Controlman 3 and 2, Rate Training Manual, NAVPERS 10571-F, Bureau of Naval Personnel, Revised 1969.

11. Disaster Control (Ashore and Afloat), NAVPERS 10899-B, Bureau of Naval Personnel, Revised 1968.

12. Shipboard Damage Control, NWIP 50-3, Department of the Navy, 1971.

13. Instruction Manual for Oxygen Breathing Apparatus, Navy Type A-1, A-2 and A-3, NAVSHIPS 393 0573, Mine Safety Appliances Co. Pt. No. 74004, Undated.

14. Technical Manual, Type J, Operation and Service Manual with Illustrated Parts List for Fire Extinguishing System, Twin Agent (AFFF and PKP), NAVSHIPS 0993-023-6010, Naval Ship Systems Command, 1973.

15. The Bluejackets' Manual, Naval Institute Press, 1973.

16. Fireman, Rate Training Manual, NAVPERS 10520, Bureau of Naval Personnel, 1971.

17. Curriculum for K-780-402 Advanced Damage Control, Fleet Training Center, San Diego, California.

18. Curriculum for K-780-404 Material Conditions of Readiness, Investigation, and Shoring, Fleet Training Center, San Diego, California.

19. Curriculum for K-780-2004 General Damage Control, Fleet Training Center, San Diego, California.

20. Damage Control Assistant's Handbook, Naval Schools Command, San Francisco, California, 1970.

21. Division Damage Control Petty Officer Handbook, Fleet Training Center, San Diego, California.

22. Hoyt, W. G. and Butler, A. K. Computer Integrated Instruction General Damage Control Learning Objectives. TN-01, prepared by System Development Corporation for Navy Personnel Research and Development Center, San Diego, California, 29 March 1974.

23. Hoyt, W. G. and Butler, A. K. Computer Integrated Instruction General Damage Control Learning Objectives. TN-01A, prepared by System Development Corporation for Navy Personnel Research and Development Center, San Diego, California, 15 July 1974.

24. Hoyt, W. G. and Butler, A. K. Computer Integrated Instruction General Damage Control Module Specifications. TN-02, prepared by System Development Corporation for Navy Personnel Research and Development Center, San Diego, California, 30 April 1974.

25. Hoyt, W. G., Butler, A. K., Hayward, C. D. Shipboard Computer Integrated Instruction in General Damage Control: Development Phase. NPRDC TR-76-XX, Navy Personnel Research and Development Center, San Diego, California, August 1975.

"CMI--NAVY STYLE"

Worth Scanland, Ph.D.
Staff, Naval Education & Training Command

The Navy has placed into operation, following several years of research and development, a computer application to instruction of which it is rather proud, and it is the purpose of this paper to describe this system and its major characteristics. But first we will speak to some instructional philosophy by way of introducing CMI, Navy style.

Because we will be discussing one aspect of instructional technology, we should have a commonly acceptable definition of that term. In 1970 a Commission of Instructional Technology made its report to the President and the Congress, in which the following definition was offered: "Instructional technology is a systematic way of designing, carrying out and evaluating the total process of learning and teaching in terms of specific objectives, based on research in human learning and communications, and employing a combination of human and nonhuman resources to bring about more effective instruction." Considering the source and the global sense of that definition, the Navy finds it very satisfactory.

For purposes of this discussion we will suggest that there are two major dimensions to the application of instructional technology. The first of these is the technology involved in the structuring of a course or program of instruction...in the military training commands this is by common agreement called "instructional systems development." The second dimension is that activity which makes available to the student the media for effecting the desired behavioral changes, or put another way, allows him to achieve his learning objectives. We might refer to this dimension as the "delivery system", and there are, broadly speaking, two major categories of these delivery systems. One we could call "conventional group" to describe the kind of instructional delivery system with which most of us can associate when we recall our own personal school experiences. This, of course, is the system where thirty or so students sit in a classroom and listen to the facts which the teacher has decided are most helpful for them to know. The other we can call "individualized instruction", which I will attempt to define for us. But before we do that, let us look at the delivery subsystems, if you will, that could fall within the general classification of methods by which we deliver instruction. They would include text books, lectures, discussion, tutoring, demonstration, laboratory, experimentation, programmed instruction, and computer assisted instruction, to name a few. Now the point we are trying to make here is that these subsystems fit into any major system of instructional delivery, and that includes both conventional group and individualized instruction. If

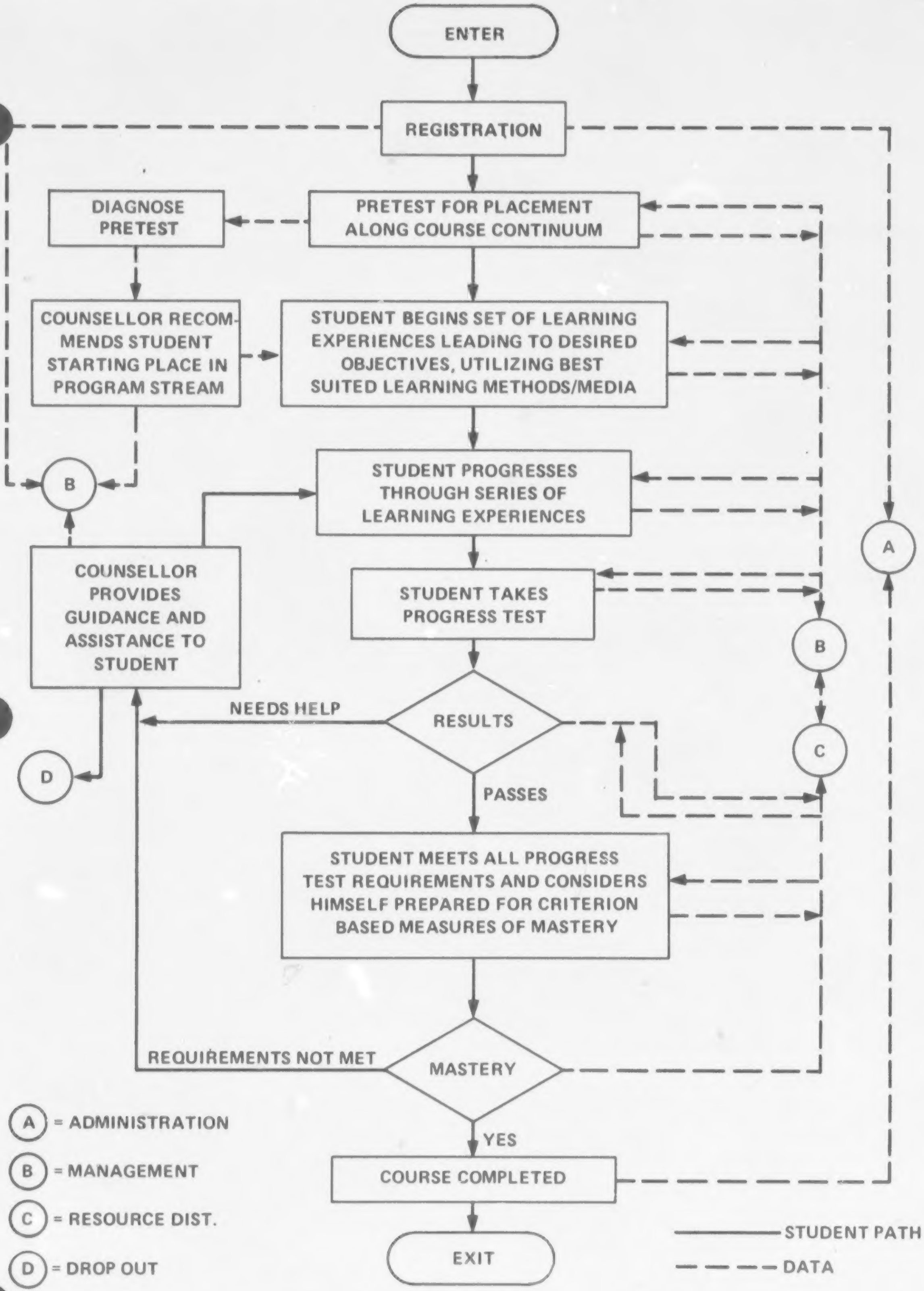

ENTER
REGISTRATION
DIAGNOSE PRETEST
PRETEST FOR PLACEMENT ALONG COURSE CONTINUUM
COUNSELLOR RECOMMENDS STUDENT STARTING PLACE IN PROGRAM STREAM
STUDENT BEGINS SET OF LEARNING EXPERIENCES LEADING TO DESIRED OBJECTIVES, UTILIZING BEST SUITED LEARNING METHODS/MEDIA
B
STUDENT PROGRESSES THROUGH SERIES OF LEARNING EXPERIENCES
A
COUNSELLOR PROVIDES GUIDANCE AND ASSISTANCE TO STUDENT
STUDENT TAKES PROGRESS TEST
B
NEEDS HELP
RESULTS
C
D
PASSES
STUDENT MEETS ALL PROGRESS TEST REQUIREMENTS AND CONSIDERS HIMSELF PREPARED FOR CRITERION BASED MEASURES OF MASTERY
REQUIREMENTS NOT MET
MASTERY
A = ADMINISTRATION
YES
B = MANAGEMENT
COURSE COMPLETED
C = RESOURCE DIST.
STUDENT PATH
EXIT
D = DROP OUT
DATA

FIGURE 1

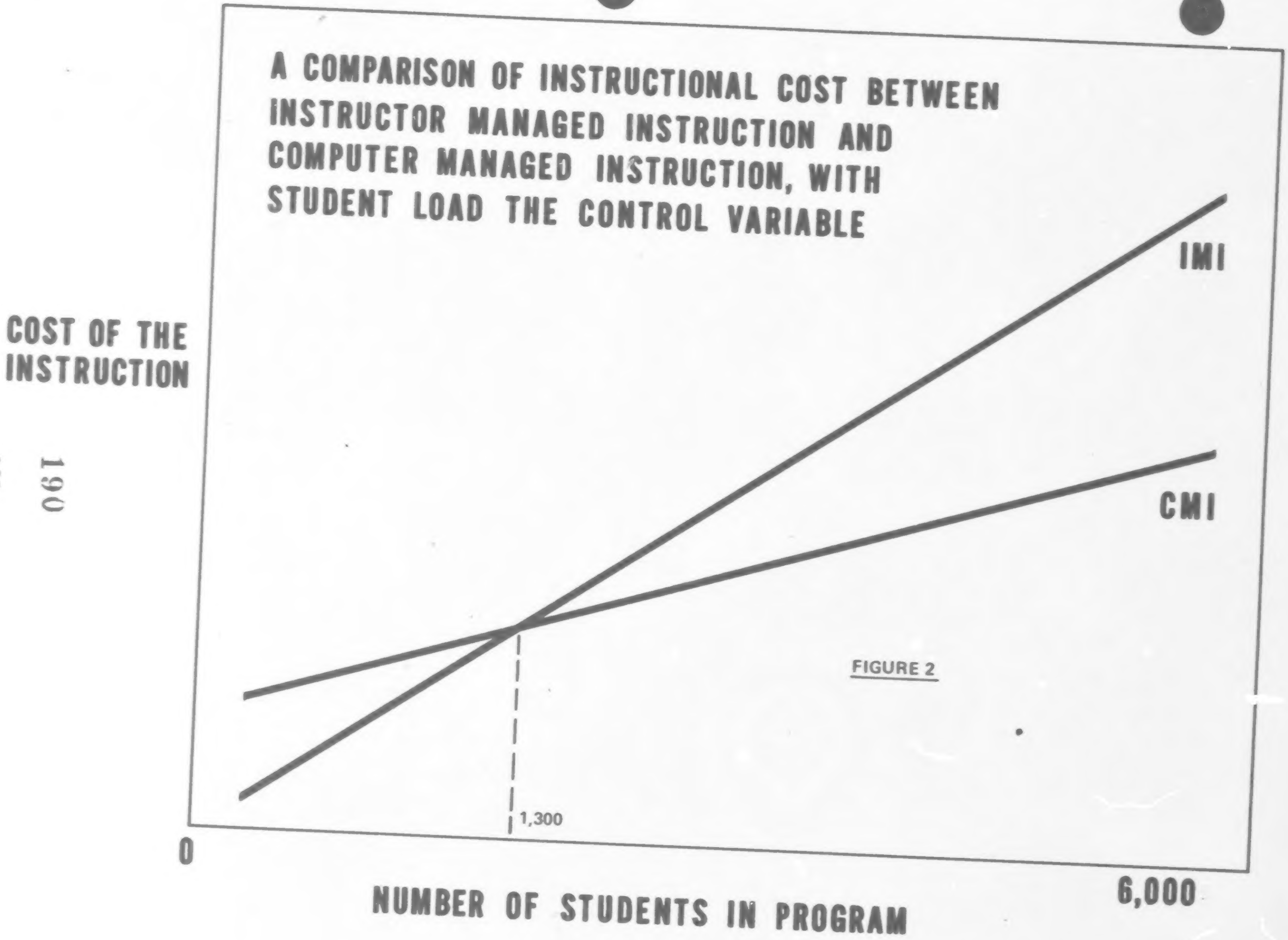
that be the case, what makes individualized instruction so different that it warrants attention? Individualized instruction uses the same subsystems as other kinds of instruction, but is different because it attends to a certain very, very important human characteristic, called individual differences.

At least in the Navy, individualized instruction attends to individual goals, to individual entry skill levels, to individual learning rates, to idiosyncratic learning methods, and to criterion rather than normative referenced measures for determining mastery of objectives by the student. Now if we attend to all these individual differences we make instruction truly humanistic, but we also make it very difficult for the instructor to manage...he no longer has thirty or so students to whom he can give out a common assignment, or lecture to, or administer a test for normative discrimination. Now he must attend to thirty (or some number of) students each of whom is in effect a class by himself.

To give an example of the kinds of communications and data exchanges required to cope with a well individualized instructional program, let us examine the flow diagram shown in Figure 1, which is intended to show a student entering an individualized course and proceeding down through the network of activities. The solid lines represent the path of the student through the program, while the dotted lines represent the information and data flow. As one can see, at every step of the process there is some kind of information or data flowing back and forth between various elements of the system. If an instructor is managing this operation, one can appreciate the kind of demand which is placed upon his ability to listen, diagnose, prescribe, test, counsel and administer. In fact, it can quickly become overwhelming if the number of students passes some critical level, probably between ten and fifteen. If the number of students exceeds ten to fifteen, the instructor must get an assistant instructor, and so on.

Now we have arrived at the point where we can make the case for the Navy concept of individualized instructional management. There are three basic axioms upon which the Navy CMI system has been built. These are, first, that well delivered individualized instruction demands a very high level of information and data flow and processing, second, that in these times of automated data processing technology, large quantities of data justify the application of ADP equipments and techniques, and finally, that thus viewed, ADP process is intended for the management of instruction, not for its delivery, and there is a significant difference. We have implacably resisted in many well-intentioned efforts to corrupt our management system with "just a little delivery, please", and will continue to resist until our established objectives have been fully met...than we may reconsider.

When one is contemplating the problem of how much data can be manually handled before the high costs of ADP equipments are seriously considered, he usually turns to other's experiences for useful information. There was,



A COMPARISON OF INSTRUCTIONAL COST BETWEEN INSTRUCTOR MANAGED INSTRUCTION AND COMPUTER MANAGED INSTRUCTION, WITH STUDENT LOAD THE CONTROL VARIABLE

FIGURE 2

however, at the time when we were faced with these questions, no other experience upon which we could draw. No one had previously thought of managing hundreds, let alone thousands, of students in an instructional program through automated data handling techniques, so the alternative to real world experience was simulation. With the assistance of a federal group known as FEDSIM, we developed a computer program which would reasonably simulate our CMI system as it was envisioned and tested in a research mode. By altering the principal variables of student loading and instructor to student ratios and costs, we were able to arrive at the indicated conclusion that somewhere in the neighborhood of 1,300 students on the system at one time caused the "instructor managed instruction" curve to cross the computer managed instruction curve, and that thereafter as the number of students increased, the savings attributable to CMI also increased, as is seen in Figure 2. It was this argument that we presented to the Navy Department and the OSD ADP people, and which won for us the approval and funding for the acquisition of our new hardware, the purchase of which was consummated last spring. I might say parenthetically here that in the meanwhile we have renting computer time on a neighborly machine located twenty miles from our instructional site.

Figure 3 is a fairly simplified diagram intended to show the principal features of the Navy CMI system. As is usually the custom in such systems, the student is registered on the computer, which acknowledges this action by giving him a student number and directing him to a pretest to determine his entry level into the program. The student takes this pretest and feeds to an optical scanner the answer sheet he has just completed. The scanner reads the sheet and transmits the information to the computer, returning the answer sheet to the student. In an average turn-around time of about 12 seconds the computer diagnosed the answers, prescribes a learning module for the student and delivers this to him at the terminal by way of a small, high speed printer. The student then tears off his prescription and carries out the directions offered to him. At the end of that learning experience he takes a progress test and repeats the procedure with the optical scanner, iterating this process until the computer informs him that he has successfully indicated mastery of all the objectives set for him and has therefore completed the program. As one can see, this is very similar to the flow diagram shown in Figure 1 to describe the well developed individualized instructional program--what we have done here is substitute a computer program for the counsellor and the administrator. The student does all his learning from materials provided from the learning center, not from the computer. If he runs into trouble and needs help, there is someone (the Learning Center Supervisor) on hand to provide it, although usually he is able to move along entirely on his own.

The Learning Center Supervisor also uses the computer to assist him in the administration and management of his students. For example, he may wish to know whether student Jones is falling behind the predicted rate

UNITED STATES NAVY COMPUTER MANAGED INSTRUCTION

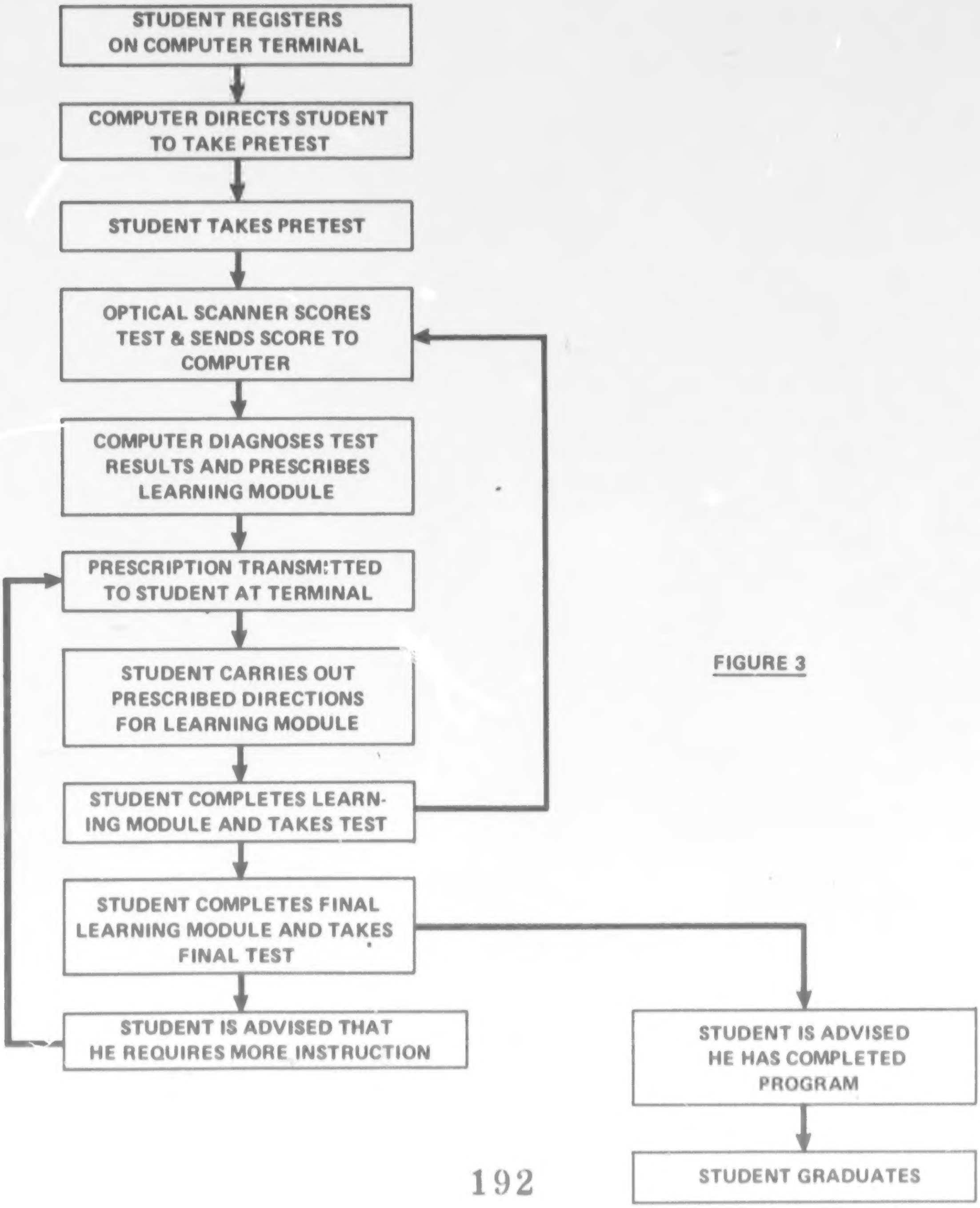

STUDENT REGISTERS ON COMPUTER TERMINAL
COMPUTER DIRECTS STUDENT TO TAKE PRETEST
STUDENT TAKES PRETEST
OPTICAL SCANNER SCORES TEST & SENDS SCORE TO COMPUTER
COMPUTER DIAGNOSES TEST RESULTS AND PRESCRIBES LEARNING MODULE
PRESCRIPTION TRANSMITTED TO STUDENT AT TERMINAL
STUDENT CARRIES OUT PRESCRIBED DIRECTIONS FOR LEARNING MODULE
STUDENT COMPLETES LEARNING MODULE AND TAKES TEST
STUDENT COMPLETES FINAL LEARNING MODULE AND TAKES FINAL TEST
STUDENT IS ADVISED THAT HE REQUIRES MORE INSTRUCTION
STUDENT IS ADVISED HE HAS COMPLETED PROGRAM
STUDENT GRADUATES

FIGURE 3

of advance, or what the current average time for course completion may be, or where student Smith is today in his progress through the program. These and other kinds of information are available to him from the computer upon inquiry. The Navy has recently purchased a large Honeywell Series 60 computer which is in process of installation at the training facility near Memphis, Tennessee. Learning centers are now in operation at both Memphis and San Diego, both sites using leased time on a Xerox Sigma 9 machine located at the Memphis State University. Plans are in process to install additional learning centers at the Great Lakes Naval Training Center and at Pensacola, to be tied in with the new computer under installation at Memphis. This installation is due for completion September, 1975.

There have been some very encouraging achievements during the early life of the Navy CMI system. There are currently about 3,000 students on the system at any one time, and this is a growth from 25 in the summer of 1973. Because of shortened student time to mastery, we have saved an estimated ten million dollars in student salaries, or at least the charging of those students' time to the training command. This has been done through the reduction of mean student time to mastery by an average of over 42 percent, and we need to make it clear here that this shortening is not because of CMI--rather it is because of the good design of the programs and the individualization of the instruction, the latter of which would have been impossible for the numbers of students we are dealing with were it not for the availability of CMI. In addition, we have seen a mean reduction of student to staff ratio of 23%, and that is truly attributable to CMI. It gives a partial answer to the restraints the Congress are placing upon us for reductions in the costs on training and the size of training staffs. And finally, this has all happened while the students' performance has improved and the attrition rate has reduced.

It is hoped that these data will give you some idea of the reasons for the enthusiasm we have for CMI, Navy Style, and the reasons we believe it is the way of the future. Speaking of the future, here are our plans for the years ahead. By 1978, barring budget cuts which remove our capability to expand, we will have 17,000 students on the system. And let us understand that were it not for the availability of CMI, we would need 1,700 more instructors to manage that number of students in an individualized mode, or we would be forced to put the students back into a lockstep mode, and that would add some 5,000 more students to the pipeline, at an annual cost of 25 million dollars. By 1980 we expect to substitute satellite communications techniques for our currently very expensive land lines for the transmission of data between the central computer and outlying stations, and perhaps most exciting of all, we will be managing the instruction of students on ships at sea and remote overseas bases by our CMI system delivered by these same satellites. That will truly be CMI, Navy Style!

LEGAL POLICIES CONCERNING TESTING
AND THEIR EFFECT ON MILITARY PERSONNEL

by
Darryl Lang, U.S. Coast Guard

INTRODUCTION

The great increase in the use of psychological tests in the past twenty years by both private and public employers has brought to public attention the question of the effectiveness and fairness of these tests. The employer is responsible for seeing that the test he or she uses fits the needs of the organization and equally important---that the test does not eliminate any man/woman or groups of people from competing equally for a job for which they are qualified. This paper concerns the laws, policies, and legal events related to the use of tests and the responsibility of the employer to demonstrate that the test is a valid predictor of on-the-job performance.

First, for those of you who missed or do not recall the paper presented at MTA last year concerning public policies and the use of tests, I will present a review of the legal developments and policies.[8] Secondly, a detailed account of two recent and relevant court cases will be presented. And lastly, some comments on how the laws and policies pertain to military testing and some suggestions as to what role the military should play in the development of guidelines concerning the testing issue.

REVIEW OF LEGAL DEVELOPMENTS AND POLICIES

In July of 1964, the Civil Rights Act became public law [10]. Title VII, commonly known as the Tower Amendment, states that it is unlawful for a private employer to give or act upon the results of an ability test that is designed, intended, or used to discriminate because of race, color, religion, sex, or national origin.

One year later, President Johnson issued an Executive Order (11246) which prohibits discrimination in Federal employment on the basis of race, creed, color, or national origin [6]. The Order called for the establishment of programs throughout the executive departments and agencies to carry out the objectives of equal employment in the Federal government. The Civil Service Commission was given the role of issuing regulations, order, and instructions to insure that the goals of the programs were carried out and to supervise the programs for civilian employees and applicants. Some of you may have noticed that this Executive Order does not mention discrimination on the basis of sex. It was not until 1967 that the late President amended this Order to read---discrimination because of race, color, religion, sex, or national origin [7].

The Civil Rights Act of 1964 also established the Equal Employment Opportunity Commission (EEOC) whose goal would be to insure employers' compliance with the intent of the Act. In order to clarify and interpret Title VII, "Guidelines on Employment Testing Procedures" were published by the EEOC [4]. However, the generality and vagueness of some of the terms and guides led to the publication of another set of guidelines four years later [5]. The 1970 Guidelines clarified and expanded the issues raised in the '66 publication.

At this point, I would like to outline the contents of these 1970 Guidelines due to the fact that in nearly every court case I have read, the Guidelines are discussed and used to interpret Title VII. The Guidelines define discrimination as the use of any test which adversely affects hiring, promotion, and the transfer of peoples protected by Title VII unless the test has been validated and evidences a high degree of utility. According to the Guidelines a test's validity should be based on empirical data demonstrating that the test is predictive of or highly correlated with elements of on-the-job behavior. The American Psychologial Association's "Standards for Educational and Psychological Tests and Manuals" first published in 1966, is referred to in defining the different types of validity. The accepted procedure for validation, according to the Guidelines, is criterion-related validity. Construct or content validity may be appropriate techniques when criterion-related validity is not technically feasible. Construct and content validity should be accompanied by information which demonstrates the relevance of the content or construct to the job.

The Guidelines explain the minimum standards for validation including such topics as minority representation in predictive and concurrent validity studies, the standards for determining what criterion measures to use, the level of significant relationships between the test and at least one relevant criterion measure---differential validity studies are to be conducted by the employer and the data is to be reported for minority and non-minority groups.

The Guidelines also state that under no circumstance should the test's author, test names or descriptors, promotional literature, or the like be accepted in lieu of empirical validity studies. The term " technical feasibility", the requirement for differential validity, and the reliance on criterion-related validity has raised much controversy as will be exemplified in the court cases to follow.

The most frequently cited court case concerning testing and discrimination in employment is the Griggs versus Duke Power Company---the only Supreme Court decision on the use of employment tests as being discriminatory under Title VII[16]. Black employees of the Duke Power Company claimed that the hiring requirements of a high school diploma and a passing score on an IQ test rendered a disproportionate number of blacks ineligible for hire. The U.S. District Court in North Carolina dismissed the case---the dismissal was appealed and in absence of any discriminatory intent on the part of the employer, the appeals court decided that the hiring requirements were allowable under the Civil Rights Act. The case was sent to the Supreme Court for review and the Supreme Court reversed the remand of the Appeals Court. Using the EEOC Guidelines as an interpretation of the intent of Title VII, the Supreme Court decided the requirement of a high school degree and a passing score on an IQ test were unfair employment practices because: (1) neither standard was related to job performance, and (2) both requirements operated to disqualify blacks from being hired at a higher rate than whites. The Court stated that " ...Congress has commanded...that any tests used must measure the person for the job and not the person in the abstract" [16].

In 1972, the Equal Employment Opportunity Act was approved[11]. This Act expanded the coverage and enforcement power of Title VII and the Civil Rights Act. The EEOC was given the power to prevent persons from engaging in unlawful employment practices by reviewing all relevant cases, and by mediating the disputes before civil action is taken. The EEOC would be responsible for cases involving the government, governmental agency, or political subdivisions. The Civil Service Commission shall have the authority to insure that the intent of the Act is carried out in the federal government. The Act provided for the establishment of the Equal Employment Opportunity Coordinating Council composed of the Secretary of Labor, Chairman of EEOC, Attorney General, Chairman of Civil Service Commission, and the Chairman of the Civil Rights Commission. The Council shall have the responsibility for developing and implementing practices and policies for equal employment. Last year at this time the Council was in the process of writing a draft of the "Uniform Guidelines on Employment Selection Procedures", the latest report is that the new Guidelines will be out in a few months proposing minimum standards for testing procedures.

CASE I: JOSEPH MOODY v. ALBEMARLE PAPER COMPANY

The courts have been faced with the task of interpreting Title VII in a large number of civil rights disputes claiming employment discrimination of the basis of test results. Most of the cases involve private employers---a case recently reviewed by the Supreme Court, Moody v. Albemarle Paper Company began nearly ten years ago when black employees at the company filed suit claiming that their employer's hiring and promotional policies violated Title VII [17]. Applicants were required to have a high school diploma and a passing score on two ability tests. On the eve of the trial and shortly after the Griggs-Duke Power decision, the company hired a psychologist to carry out a concurrent validity study. He compared ten job groupings and supervisory ratings with test scores and found that for both tests only two job groupings produced significant correlations. The District Court, however, felt that the company had undergone sufficient validation studies. The court did state that the high school education requirement was unlawful in that the aptitude tests were sufficient to measure mental ability of the applicants. Thus, the District Court denied preliminary injunction. The case was appealed and the Appeals Court reversed the decision claiming that the tests should have been enjoined because,based on the EEOC Guidelines, the test was not properly validated. The case went to the Supreme Court for review. The Supreme Court agreed with the Court of Appeals and based on the Guidelines found that the validation study was defective for a number of reasons: (1) the few significant correlations between test scores and ratings were not sufficient proof of validity, (2) there had not been a study of skills or attributes of the job, (3) the supervisory ratings were not acquired in an acceptable manner, and (4) the study involved job-experienced, white workers, very different from the characteristics of the applicants. Thus, the Supreme Court remanded that the District Court make a decision in the case and review the company's reorganization and new hiring requirements implemented subsequent to the law suit.

The Supreme Court's reliance on the EEOC Guidelines was not agreed upon by all member of the court. One judge claimed that there was too much reliance on the Guidelines and they should be related to the testimony of an expert in the field. Another judge felt that an employer should be required to show that the test is fairly related to job skills and work characteristics, but that " rigid application of the EEOC Guidelines will leave the employer little choice save an impossibly expensive and complex validation study, but to engage in a subjective quota system of employment. This, of course, is far from the intent of Title VII" [17].

CASE II: DOUGLAS ET AL. v. HAMPTON

Another important case I would like to mention is the Douglas versus Hampton case involving the fairness of the Federal Service Entrance Exam (FSEE) given by the Civil Service Commission to select college graduates for federal employment [14]. In 1971, eight black college graduates filed suits claiming that the use of the FSEE was unlawful in that it rejected a disproportionate number of minority applicants and was not predictive of job performance. The District Court denied an injunction based on the claim that the plaintiffs failed to show that the FSEE had a racially disportionate impact and that the Civil Service Commission's construct validity technique had insufficiently demonstrated that the FSEE was a reasonable measure of job performance. The Court remanded the Commission to process the complaint according to the new procedures issued by the Commission subsequent to the lawsuit [2].

In February of this year. the Court of Appeals reversed the ruling of the District Court. According to the Appeals Court. the District Court's decision to deny preliminary injunction was based on the erroneous premise that the procedure of construct validity is acceptable without demonstrating that empirical (or criterion-related) validity is technically feasible. Thus. the Appeals Court remanded to the District Court to look into the technical feasibility of the Civil Service Commission carrying out an empirical validity study. The Appeals Court's reliance on the EEOC Guidelines in reaching this decision was not looked upon favorably by the Civil Service Commission. In a Civil Service Commission Operations memo dated April 30. the Commission claimed that the court did not consider the standards and procedures for test validation promulgated by the Civil Service Commission or the fact the Commission. not the EEOC, is the agency responsible for enforcement of equal employment opportunity in the federal service as stated in the Equal Employment Opportunity Act of 1972 [3]. The Commission also stated that the EEOC Guidelines lack both judicial sanction and the support of the psychological profession. The Commission feels that this case is perhaps "...the most significant Civil Service question to be raised in court since the creation of the Commission"[3].

Since the filing of the lawsuit. the Civil Service Commission no longer uses the FSEE. but has developed another entrance exam. Professional and Administrative Career Exam (PACE) which was developed using the job-analysis procedure proposed by the Commission but not accepted in the EEOC Guidelines unless a criterion-related validity study is technically infeasible.

Two months ago the Civil Rights Commission released the fifth volume of a report entitled "The Federal Civil Rights Enforcement Effort---74" [15]. This volume analyzed the equal employment practices and programs of the Civil Service Commission. In regard to the selection procedures for entrance into the federal service. the Civil Rights Commission critized the methods of selection claiming that the construct validity procedure advocated by the Civil Service Commission was not sufficient to demonstrate that the tests fairly measured how one would perform on the job. In referring to the EEOC Guidelines, the Civil Rights Commission indicated that the development of the new federal service selection test. the PACE. had not been properly validated and there was no evidence that a particular score on any of the PACE tests actually predicts job performance in any job. The report concluded that "...the Commission screens out and ranks thousands of applicants annually on the basis of examination...which have not been demonstrated empirically to predict job performance ... Since the Commission has not demonstrated the validity of its procedures, it has no scientific basis on which to assert that candidates, at least within ranges of scores, are not equally qualified" [15].

CONCLUSIONS AND RECOMMENDATIONS

The critical issue in testing and fair employment appears to be not only whether a test should be predictive of how a person will perform on a job, but what validation technique will accurately measure this relationship. In the courts, the Civil Service Commission has the task of proving their validation technique is sufficient to demonstrate that their test results accurately reflect the person best qualified for the job. The psychological profession, other than reference to the American Psychological Association's Standards in the EEOC Guidelines, has not played a prominent role in the shaping of federal legislation and court rulings [9]. The testing issues are far from being resolved.

How do these civil cases, guidelines, laws, and opinions affect military testing? The nearly three million men and women tested each year for selection into the services are protected under Title VII. The enlisted men and women classified by ability tests certainly do have the right to dispute any discriminatory classification practices. We are all actively conducting research in a number of differenct personnel areas, including the validation of the new forms of the Armed Services Vocational Aptitude Battery (ASVAB). To my knowledge there are no specific instructions promulgated by any of the services as to what guidelines to use for validating and using tests. Civil cases have been filed in court questioning the different cutting scores for selecting men and women into the service, e.g. Parise v. Schlesinger. Three years ago a report released by the Armed Forces Task Force on Military Justice mentioned cultural bias in military selection and classification procedures [12]. The report concludes, after briefly mentioning some unreferenced studies, that cultural bias in the tests is minimal, tests tend to overpredict performance of minority persons and the aptitude tests are valid for Mental Categories I-III, but not necessarily for Category IV personnel. A study released by the Navy a few months ago affirms that the relationship between their selection composites on the BTB and technical school performance are different for whites and blacks [13]. Of the twenty-five schools studied, fourteen of them showed that the black test scores under predict their school performance.

Up to now we have not been involved in the formulation of guides, laws, and policies on testing. Our personnel practices which rely heavily on psychological testing will perhaps be under scrutiny similar to those of the private employers and the Civil Service Commission. Thus, I believe that it is our professional responsibility to become involved in the formulation of guidelines on the construction, validation, and use of tests based on recent developments, further research, and our expertise in the field.

REFERENCES

[1] American Psychological Association. *Standards for Educational and Psychological Tests*, Washington: D.C., 1974 (revision of *Standards for Educational and Psychological Tests and Manuals*, 1966).

[2] Civil Service Commission. Testing and Employee Standards Instruction. 37 *Federal Register*, 21552 (1972).

[3] Civil Service Commisssion Memo, April 30, 1975. *Recent Court Decisions on Civil Service Commission Employment Testing*.

[4] Equal Employment **Opportunity** Commission. Guidelines on Employment Testing Procedures. *Federal Register*, August 24, 1966.

[5] Equal Employment Opportunity Commission. Part 1607-Guidelines on Employee Selection Procedures. *Federal Register*. Vol 35, August 1, 1970.

[6] Executive Order 11246. Equal Employment Opportunity. *Federal Register*. September 28, 1965, 12319.

[7] Executive Order 11375 Equal Employment Opportunity. *Federal Register*. October 17, 1967, 14303.

[ ] Lewiski, Raymond L. "Public Policy Concerning the Use of Tests" paper presented at the 16th Annual Conference of the Military Testing Association, October 1974.

[9] Fincher, Cameron. "Psychological Testing and the Public", *American Psychologist*, June, 1973, p. 489-497.

[10] Public Law 88-352, as approved by the President on July 2, 1964; *Civil Rights Act of 1964 with Explanation*, Title VII, p. 253-266.

[11] Public Law 92-261; as approved by the President on March 24, 1972; *Equal Employment Opportunity Act of 1972*, p. 103-113.

[12] Report of the Task Force of the Administration of Military Justice in the Armed Forces, November 30, 1972, Vol II, "Testing, Job Assignments and Promotion", p. 11-19.

REFERENCES (cont'd)

[13] Thomas, Patricia J. Racial Differences in the Predication of Class "A" School Grades. San Diego: Navy Personnel Research and Development Center, TR-39-75.

[14] U.S. Appeals Cout, D.C., *Douglas et al. versus Hampton*. No. 72-1376, February 27, 1975.

[15] U. S. Commission on Civil Rights. *The Federal Civil Rights Enforcement Effort-1974,* Vol. V, July, 1975.

[16] U.S. Supreme Court, *Griggs et al. versus Duke Power Company*. No. 124-October Term, 1970(March8, 1971).

[17] U. S. Supreme Court Reports, Lawyer Edition. *Albemarle Paper Company versus Joseph Moody,* Vol. 45, L.Ed 2d, No. 2 August 8, 1975, p. 280-318.

ADDITIONAL REFERENCES

Cooper, George and Richard B. Sobol. " Seniority and Testing Under Fair Employment Laws: A General Approach to Objective Criteria of Hiring and Promotion", *Harvard Law Journal*, 82: 1598-1679, June 1969.

Guion, Robert M. "Open a New **Window:** Validities and Values in Psychological Measurement", *American Psychologist,* May, 1974, p. 287-296.

Koenig, Peter, "They Just Changed the Rules on How to Get Ahead", *Psychology Today*, June, 1974.

Miner, John B. "Psychological Testing and Fair Employment Practices, A Testing Program that Does not Discriminate", *Personnel Psychology*, 1974, vol. 27, p. 49-62.

Roth, David C. and Philip C. Stahl, "Application of EEOC Guidelines to Employment Test Validation: A Uniform Standard for Both Public and Private Employees", *George Washington Law Review,* 41, 505-537, 1972-73.

Sharf, James C. "How Validated Testing Eases the Pressures of Minority Recruitment". *Personnel*, May-June , 1975, p. 53-59.

ONE MOS PROFICIENCY TEST TO COVER ALL SKILL LEVELS

by

Herman A. Mahnen

The use of a single global test for the concurrent assessment of varying levels of some characteristic is neither new nor original. The 1905 Simon-Binet Scale, which consisted of 30 items ranging in difficulty from those suitable only for the classification of very low-grade idiots to those intended for children in the upper elementary grades, is just one example of such a global test.

This paper reports on the results of the field trials of an evaluation test which covered all three skill levels of a Military Occupational Speciality (MOS 36C Lineman). The primary purpose of the field trial was to ascertain the feasibility of using multi-skill level tests in the U. S. Army Enlisted Evaluation System (USAEES). If feasibility could be established, then multi-skill level evaluation tests could be introduced - first, on a limited basis and gradually extended to include most of the more than 400 MOS in the U. S. Army. The advantages of multi-skill level tests are obvious; they offer greater parsimony of both time and money.

The Army requires periodic evaluation of the relative competence of enlisted soldiers, with more than 17 months of service, in relation to the requirements of their MOS. Two instruments are usually used in these evaluations: an MOS evaluation test and an Enlisted Efficiency Report. The evaluation test is a paper-and-pencil test made up of 125 multiple-choice items. The Enlisted Efficiency Report (EER) is a supervisory evaluation of how well the soldier is performing the duties of his present assignment, i.e., it is the supervisor's subjective appraisal of the soldier's performance. For several MOS, a performance test is used in addition to the evaluation test and EER. These performance tests, which include typing, stenography, etc., are used in those MOS codes where performance testing is determined to be relevant and feasible.

Since the establishment of the USAEES in 1958, soldiers have been evaluated on an annual basis. In general, MOS evaluation tests are also revised on an annual basis. There is usually an overlap of some 25 percent, or even more, of the test items in the evaluation tests for adjacent skill levels in an MOS. Occasionally, a test has been used for soldiers in two skill levels - usually the 10 and 20 - or elementary skill levels in an MOS. It has long been Army policy, however, that a separate evaluation test be developed for each skill level in an MOS, i.e., for each four-character MOS Code.

Supplanting MOS tests which cover only one skill level with multi-skill level tests offers several advantages:

1. Cost reductions should be realized in the MOS evaluation program by decreasing the number of tests to be developed.

2. Savings in administration time since it would obviously require less time to administer one test covering three skill levels than three separate tests each covering one skill level.

3. Tests covering two or more skill levels could provide the means of assessing the soldier's competence at the next higher skill level as a prerequisite for promotion.

There are a number of psychometric problems which arise from tests designed to measure specific behavior at differing skill levels. Such things as does the test have adequate reliability at each skill level. Does the test clearly delineate each skill level? Moreover, to what extent can specific test items be assigned to the separate and distinct skill levels? What is the degree to which specific test items are applicable to certain skill levels?

It was decided that the multi-skill level testing approach would be most feasible with a homogeneous MOS test population. Accordingly, 36C, Lineman was selected for the field trials. This MOS consists of three skill levels and has a global MOS description as follows: supervises, installs, performs organizational maintenance, and rehabilitates single and multi-conductor aerial and buried communication systems.

Most soldiers holding MOS 36C are assigned to small units. They are found, in the main, in cable/wire installations sections (or teams) of signal battalions. Typically, they serve as wiremen, senior wiremen, or as section or team chiefs who supervise wiremen. There are, however, certain unusual duty position which are sparsely populated, e.g., the main distribution frame specialist of the command signal operations company.

The experimental multi-skill level test was developed by the Southeastern Signal School. It covered 14 major areas and included 52 items at the 20 skill level, 52 items at the 40 skill level and 39 items at the 50 skill level for a total of 143 items. Subject-matter specialists at the Army Signal School developed all of the test items which were designed to sample the 14 major duty areas and related subareas in the test plan. They designated each item in the test as being appropriate for a specific skill level. All of the items were randomly arranged in the test booklet. Eleven of the 14 major duty areas contained one or more items designated for each of the three skill levels; the remaining three major duty areas contained questions for the 20 and 40 skill levels only.

There were 328 soldiers in the total sample which was distributed by skill level as follows: Skill Level 20, 98 soldiers; Skill Level 40, 223 soldiers; and Skill Level 50, 7 soldiers. It is emphasized that the very small number of soldiers at Skill Level 50 requires some degree of caution in interpreting the findings applied to this skill level.

The field trials of the multi-skill level test were administered at five Army installations. These installations were selected because they had relatively large populations of soldiers in MOS 36C. The installations were Ft. Hood, TX., Ft. Riley, KS, Ft. Gordon, GA, Ft. Bragg, NC, and Ft. Lewis, WA.

Results

As previously mentioned, the study was concerned with two major psychometric issues: (1) the statistical functioning of the total test and subtests at the different skill levels, and (2) the problem of identifying specific items as appropriate for specific skill levels.

The test results contained in Table 1 bear most directly on the first issue, i.e, the issue relating to the discrimination between soldiers at the different skill levels, discrimination among soldiers at each skill level, and test/subtest reliability.

Table 1 confirms that the test accurately discriminated between soldiers at the different skill levels. Mean scores on the total test increased from 62.3 for soldiers at Skill Level 2 to 84.9 for Skill Level 4 soldiers and then to 104.4 for soldiers at Skill Level 5. The increase in mean scores is in the right direction for all skill levels.

This fact is borne out more graphically in Figure 1. Here, it can be seen that the increase in mean scores across skill levels is essentially a straight line progression. This pattern holds not only for the test as a whole, but also for each of the skill level subtests.

The same straight line progression for the subtests is evident in Figure 2. In order to adjust for the varying number of items in the different subtests, the mean percentage of soldiers obtaining the correct answer (item difficulty values) rather than raw score mean has been used in plotting the lines in Figure 2. The pattern of progression of scores in Figure 2 clearly shows that this linear progression is the same whether based upon the test as a whole or one of the skill level subtests. In each case, Skill Level 2 soldiers

Table 1

Overall Multi-Skill Level Test Score Results (36C-20/40/50)

All Skill Level Soldiers Combined (N=328)

NR of Questions	Skill Level Questions	Mean Score ($\bar{X}$)	Standard Deviation	$\bar{p}$*	r_{xx}**	r_{xy}**
143	20,40,50	78.61	13.62	.54	.92	.41
20 Skill Level Soldiers Only (N=98)						
143	20,40,50	62.29	14.29	.43	.86	.36
52	20	24.72	5.97	.47	.71	.37
51	40	20.68	5.67	.41	.68	.37
40	50	16.92	4.41	.42	.60	.35
40 Skill Level Soldiers Only (N=223)						
143	20,40,50	84.89	15.28	.60	.89	.39
52	20	33.39	6.17	.64	.77	.40
51	40	28.41	6.01	.56	.73	.39
40	50	22.97	5.29	.58	.75	.37
50 Skill Level Soldiers Only (N=7)						
143	20,40,50	104.42	19.74	.73	.95	.37
52	20	39.86	6.01	.76	.85	.33
51	40	37.00	7.94	.72	.90	.42
40	50	27.57	4.87	.69	.77	.33

*Difficulty value or percentage of total test questions correct.

**Reliability index (correlation value showing degree to which soldiers would get the same score if retested).

***Discrimination index (correlation value showing degree to which the best soldiers get the question correct vs the poorest scoring soldiers getting the question correct. If equal numbers of the two groups get a question correct, the correlation value would be .00).

Figure 1

Plotting of Total Scores by Skill Level

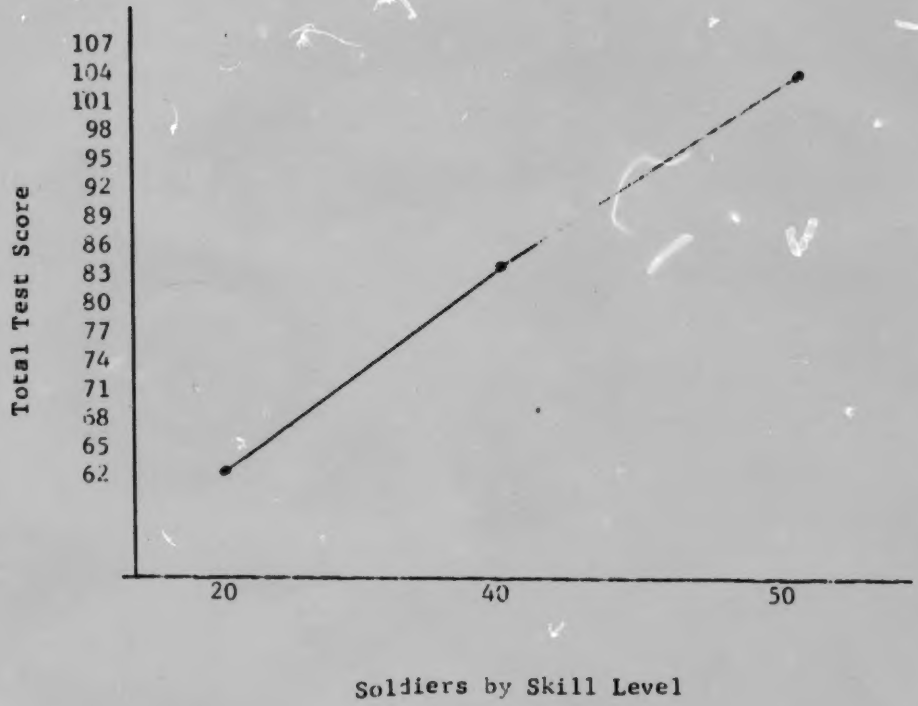
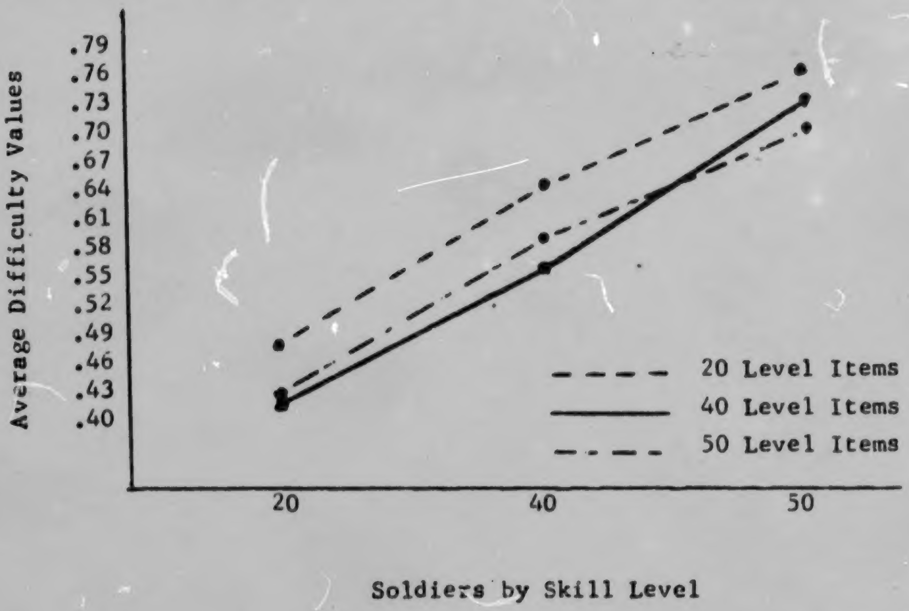


Figure 2

Plotting of Average Difficulty Values by Skill Level



scored the lowest; Skill Level 4 soldiers scored in between; and Skill Level 5 soldiers scored the highest. All of the group differences, both means and standard deviation, were statistically significant at the .05 level or better. For all groups of soldiers, scores on the Skill Level 4 subtests were lower than scores on the Skill Level 2 subtests. But contrary to what might have been expected, the Skill Level 5 subtest was easier than the Skill Level 4 subtest for both Skill Level 2 and Skill Level 5 soldiers; that is, they scored relatively higher on the Skill Level 5 than the Skill Level 4 subtest. The Skill Level 5 subtest contained items concerning managerial and general supervisory problems. These items were answered correctly more often by soldiers from all three skill levels than were the technical knowledge items.

In addition to the clear differentiation between soldiers at the three skill levels, the test as a whole, as well as the three skill level subtests discriminated rather well among soldiers at each skill level. The point-biserial correlations which are presented in Table 1 are used as indices of discrimination.

These values indicate that not only the test as a whole discriminated among soldiers at each skill level, but also that each skill level subtest had comparable discrimination power. Thus, the Skill Level 2 subtest provided as much discrimination for Skill Level 4 as among Skill Level 2 soldiers. Conversely, all skill level subtests were about equal in their power to discriminate among soldiers at each of the skill levels.

The Enlisted Evaluation Center has adopted a minimum statistical standard of .70 for reliability of MOS tests. All of the reliability coefficients shown in Table 1 exceeded this standard of reliability for each of the three skill levels. In addition, the various skill level subtests generally exceeded the minimum reliability standard. These findings indicate that establishing adequate reliability for MOS tests covering two or more skill level should not present a problem. In short, the field trials of the multi-skill level tests indicate that reliability was adequate for all of the skill levels.

The above findings support the feasibility of using a global test to assess proficiency of two or more skill levels of an MOS; however, these identical findings point out the very difficult problem of identifying specific items as appropriate for different skill levels. In testing Skill Level 2 soldiers, Skill Level 4 and Skill Level 5 items discriminated as well as Skill Level 2 items, and Skill Level 4 items were no better than Skill Level 2 and Skill Level 5 items in discriminating among Skill Level 4 soldiers. Moreover, there was considerable overlap among scores on the different skill level subtests. That is, soldiers at any skill level who scored relatively high on Skill Level 2 items also tended to score relatively high on Skill Level 4 and Skill Level 5 items, while soldiers who scored relatively low on one skill level subtest likewise tended to score relatively low on the other two skill level subtests.

This tendency toward parallel performance is reflected in the inter-correlations between subtests shown in Table 2. All of these correlations are significant and indicate a substantial relationship between performance on the various subtests.

In summation, this study addressed two major issues: (1) the psychometric functioning of the test at the different skill levels, and (2) the extent to which specific items can be assigned to specific skill levels.

In relation to the first issue, the major findings were: (1) the test did discriminate between soldiers at the different skill levels; (2) the test discriminated among soldiers at each skill level, and (3) the test exhibited adequate reliability at each skill level.

With regard to the second issue, items designated as appropriate to a particular skill level could not actually be identified with that skill level in terms of the following: (1) the extent to which the skill level subtest differentiated between skill levels; (2) the extent to which it differentiated among soldiers at each skill level; and (3) the relative difficulty of the items for the soldiers at the different skill levels.

Conclusions and Implications

Army policy regarding MOS evaluation testing has been modified since the completion of the present study. Under the new policy, evaluation tests are to be modified in scope and content and are to be called Skill Qualification Tests (SQTs). An SQT is to cover two skill levels: the soldier's present skill level and the next higher skill level. The findings of the present study indicate that it is indeed feasible to develop tests covering two or more skill levels of an MOS. These field trial findings have implications with respect to the composition of those tests. To the extent that the findings can be generalized, they suggest that many, perhaps most items on an SQT, will be items which are pertinent to both skill levels covered by the test. This means that proficiency at the higher skill level versus proficiency at the present skill level will be assessed not so much in terms of ability to perform a different set of tracks, but in terms of differential levels of proficiency on a common set of tracks.

Table 2

Inter-Correlation Between Scores Obtained by 36C20 and 36C40 Soldiers on Each Skill-Level Subtest

Skill-Level 20 Soldiers (N=98)

	Scores on Skill-Level 40 Subtest	Scores on Skill-Level 50 Subtest
Scores on Skill-Level 20 Subtest	.73	.66

Skill-Level 40 Soldiers (N=223)

	Scores on Skill-Level 20 Subtest	Scores on Skill-Level 50 Subtest
Scores on Skill-Level 40 Subtest	.72	.72

The New Army Enlisted Performance Rating Forms

by

Arthur G. Hermansen

Concepts, policy, doctrine and methodology are subject to change as part of the eternal change in political and social structures. Since an individuals status is evaluated within the context of these changes the instruments used for evaluation must reflect the standards required by the new social and political movement. From phrenology to sophisticated total assessment centers the search for more reliable evaluation instruments and methodologies continues on a daily basis. I will present here the evaluation of the revised Enlisted Evaluation Report and its new companion, the Senior Enlisted Evaluation Report as a response to new requirements.

These are the areas to be covered in this presentation:

1. Need for a new evaluation report.
2. Initial guidance given to EEC.
3. Development of the new evaluation reports.
4. Processing the new evaluation reports.
5. Field test of the new evaluation reports.
6. Significant features of the new evaluation report.

The project to revise the Enlisted Efficiency Report was instigated by General Berry, then Chief of OPO. General Berry laid down the following guidelines to be incorporated into the development of a new EER.

1. Provide for the consideration of the "whole man".
2. Simple administrative procedures.
3. Consider requirements posed by A. D. P.
4. Parallel OER report form and system to extent possible.
5. Consider eliminating "Efficiency" from the title.

The decision to develop a new EER was dictated by the need for more comprehensive and definitive information for use by:

1. Promotion Boards in selecting soldiers for promotion.
2. Commanders for more effective utilization of personnel.
3. Supervisors for purposes of counselling and training.
4. Soldiers for their own self-evaluation.

The present EER has proven its worth as a personnel management tool, but it has a number of deficiencies, chief of which is inflation. Scores have become increasingly inflated since introduction of the present EER in July 1970.

Figure 1

Pay Grade	Current EER Army Average Score
E3	87.2
E4	106.5
E5	110.2
E6	116.9
E7	120.6
E8	122.7
E9	123.7

Leniency in rating subordinates is prevalent in many organizational settings and rating forms commonly yield inflates scores; thus, the EER is not unique - but the inflation of scores has now reached a point which seriously impairs the usefulness of the EER, especially for the higher grades.

An examination of EER scores in relation to recent Promotion Board actions points up the problem as shown in Figure 2. The last CSM Board Selected about one-third of eligible E9s for CSMs, but more than 60 percent had maximum EER Weighted Average scores. Since the number of E9s with maximum EERWAs was almost twice the number selected, the EERWA score was virtually useless in the selection process. A comparable situation prevailed in the recent E8 board. About 19 percent of the eligible E8s were selected for promotion, but more than 30 percent had maximum EERWAs. The situation is somewhat better in the selection of E7s for promotion to E8, or E6s for promotion, but more than 20 percent had maximum EERWAs. The situation is somewhat better in the selection of E7s for promotion to E8, for E6s for promotion to E7, but even here inflation is a problem.

A factor in inflation is the allocation of points to descriptive ratings in the present EER. A drop of one block in any of the characteristics drops the EER score by 3.3 points, and a drop of one block in advancement potential drops the EER by 3.5 points. For an E9 to be rated as Excellent rather than Outstanding on just one of six characteristics, is to be placed in the lower half of E9 EER scores.

Figure 2

RECENT BOARD ACTION

CSM BOARD

% OF E9 WITH MAX EERWA	% OF E9 WITH MAX EER (1974)	% OF ELIGIBLES SELECTED
60.22	90.38	32.72

E9 BOARD

% OF E8 WITH MAX EERWA	% OF E8 WITH MAX EER (NOV)	% OF ELIGIBLES SELECTED
31.27	77.06	19.15

E8 BOARD

% OF E7 WITH MAX EERWA	% OF E7 WITH MAX EER (NOV)	% OF ELIGIBLES SELECTED
11.43	68.93	16.88

E7 BOARD

% OF E6 WITH MAX EERWA	% OF E6 WITH MAX EER (NOV)	% OF ELIGIBLES SELECTED
4.40	56.20	33.40

Although inflation is the major problem, the present EER has other deficiencies. The characteristics upon which the soldier is rated are general in nature and fail to provide specific information about the soldier; moreover, there is little space for comments. After the rater has indicated whether or not the soldier has First Sergeant potential or whether he supports the Equal Opportunity Program, there is room for little else. Under normal conditions, the report provides for the opinion of only one person, the rater. The only way the reviewer can evaluate the soldier is by non-concurrence with the rater. This is rarely done and only in cases of extreme differences of opinion. The aim of developing a new EER was to correct these deficiencies and to dampen inflation in EER scores.

In approaching the task of designing a new Enlisted Evaluation Report, we reviewed the technical literature on psychological rating forms, examined our experience with previous rating systems, and examined the Army Officer Evaluation Report System. We also carefully examined the Air Force and Navy enlisted rating forms. A primary initial question was whether to continue with one rating report for all pay grades. Administrative simplicity argues for one report, but difference in duty requirements between the junior and senior grades and the serious inflation at the higher grades argues against a single form. Both the Air Force and Navy use different forms for the evaluation of senior enlisted members. In the Navy, a recently implemented senior enlisted evaluation form has resulted in substantial deflation in the ratings of Chief Petty Officers.

Our decision was to develop two rating forms - a Senior Enlisted Evaluation Report (SEER) to be used in rating soldiers in Pay Grades E6 and above, and an Enlisted Evaluation Report (EER) for rating soldiers in Pay Grades E5 and below. I should point out here that we were initially undecided whether to rate E6s with the EER or the SEER. During the field trial, which I will discuss in more detail in a few minutes, we had a group of E6s rated with both forms. Our decision to use the SEER was based upon three reasons:

1. E6s rated significantly lower on the SEER than on the EER, which will assist in dampening inflation.
2. The E6 is normally in a leadership or supervisory position and his rating should include an evaluation of traits associated with leadership. And finally,
3. E6 scores will go to a centralized selection board as will those of E7s, E8s, and E9s.

The two evaluation reports are in certain respects identical to one another.

Part I (Figure 3) is common to the two forms. This part is to be completed by a Military Personnel Officer. The form is compatible with SIDPERS and the forms can be printed out by the computer. This should significantly reduce the number of errors which impede the processing of EERs.

The section of the EER to be used in the actual rating of soldiers is shown in Figure 4. The soldier is rated on ten duty performance traits and on overall duty performance and promotion potential. Rating on promotion potential is retained from the present EER. In the present form, this is the selection which gives the most discriminating ratings. We believe that this form is a significant improvement over the present EER in the following respects:

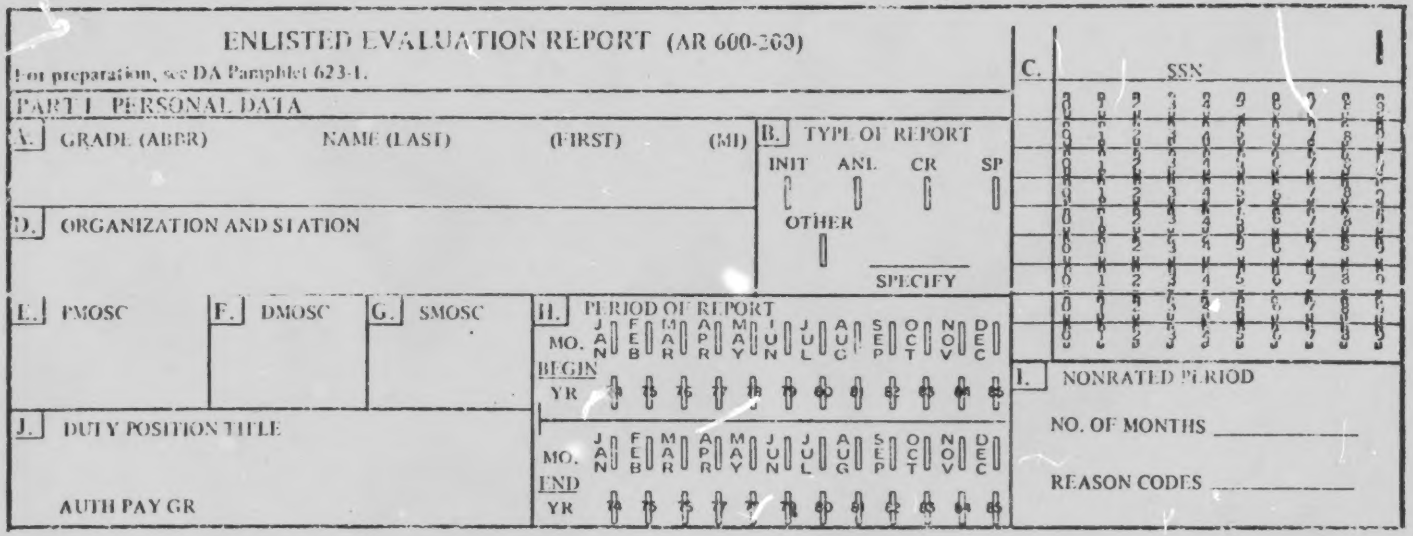
1. The traits are inclusive but specifically described. They cover the whole man and, at the same time, provide the specificity required for objective rating.
2. The scoring system is open, which enables the rater to know just how he is scoring the soldier, and the soldier to know just how his score was derived.
3. The scoring provides for finer discriminations, especially at the higher end of the rating continuum.
4. The soldier is evaluated by both a rater and an indorser. His score is an average of these two ratings.

The SEER differs from the EER mainly in Part II, Figure 5, where soldiers are actually rated. This section of the SEER is shown here. As on the EER, soldiers are rated on overall duty performance and promotion potential. But the leadership skills on the SEER have no counterpart on the EER. We believe that this is a crucial difference in the rating of junior and senior level soldiers.

Page 2 of the EER is shown in Figure 6. It includes a current development block and blocks for both rater and indorser comments. Comments are mandatory if evaluation scores given by either the rater or indorser fall above or below specified levels. Both forms include authentication blocks for the rater, indorser, soldier, and reviewer.

Page 2 of the SEER is shown in Figure 7. It is just like the EER except in the blocks in the career development section. The career development section of the SEER allows the rater and indorser to recommend E7s for First Sergeant and E8s and E9s for Command Sergeant Major.

Figure 3



ENLISTED EVALUATION REPORT (AR 600-200)

For preparation, see DA Pamphlet 623-1.

PART I PERSONAL DATA

A. GRADE (ABBR) NAME (LAST) (FIRST) (MI)

B. TYPE OF REPORT

INIT ▯ ANL ▯ CR ▯ SP ▯

OTHER ▯ ______ SPECIFY

C. SSN

0 1 2 3 4 5 6 7 8 9

D. ORGANIZATION AND STATION

E. PMOSC

F. DMOSC

G. SMOSC

H. PERIOD OF REPORT

BEGIN MO. JAN ▯ FEB ▯ MAR ▯ APR ▯ MAY ▯ JUN ▯ JUL ▯ AUG ▯ SEP ▯ OCT ▯ NOV ▯ DEC ▯

BEGIN YR 74 ▯ 75 ▯ 76 ▯ 77 ▯ 78 ▯ 79 ▯ 80 ▯ 81 ▯ 82 ▯ 83 ▯ 84 ▯ 85 ▯

END MO. JAN ▯ FEB ▯ MAR ▯ APR ▯ MAY ▯ JUN ▯ JUL ▯ AUG ▯ SEP ▯ OCT ▯ NOV ▯ DEC ▯

END YR 74 ▯ 75 ▯ 76 ▯ 77 ▯ 78 ▯ 79 ▯ 80 ▯ 81 ▯ 82 ▯ 83 ▯ 84 ▯ 85 ▯

I. NONRATED PERIOD

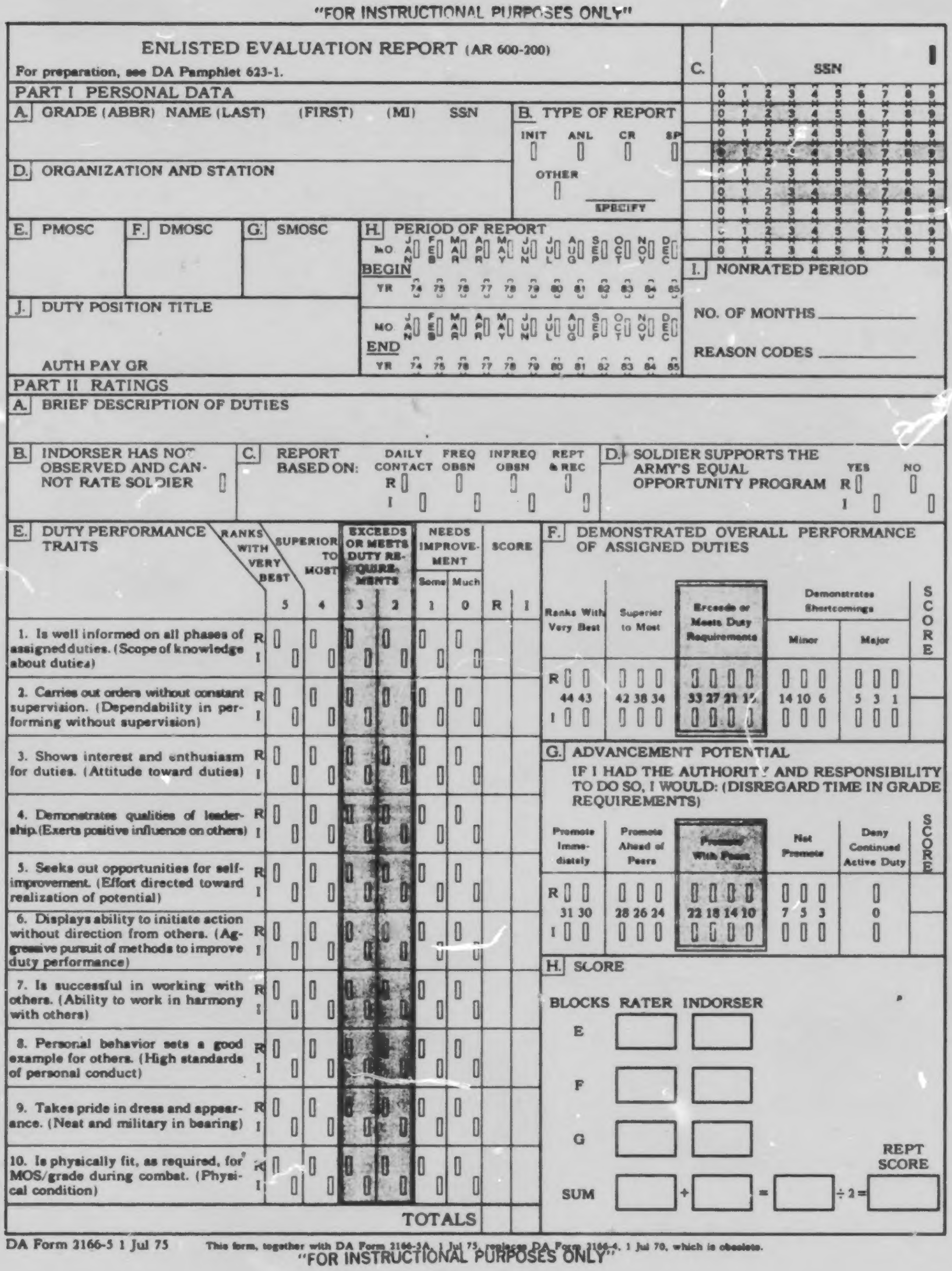
NO. OF MONTHS ______

REASON CODES ______

J. DUTY POSITION TITLE

AUTH PAY GR

Figure 4



"FOR INSTRUCTIONAL PURPOSES ONLY"

ENLISTED EVALUATION REPORT (AR 600-200)

For preparation, see DA Pamphlet 623-1.

PART I PERSONAL DATA

A. GRADE (ABBR) NAME (LAST) (FIRST) (MI) SSN

B. TYPE OF REPORT
INIT ☐ ANL ☐ CR ☐ SP ☐
OTHER ☐ ________ SPECIFY

C. SSN

0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9
0	1	2	3	4	5	6	7	8	9

D. ORGANIZATION AND STATION

E. PMOSC

F. DMOSC

G. SMOSC

H. PERIOD OF REPORT

BEGIN MO	JAN ☐	FEB ☐	MAR ☐	APR ☐	MAY ☐	JUN ☐	JUL ☐	AUG ☐	SEP ☐	OCT ☐	NOV ☐	DEC ☐
YR	74	75	76	77	78	79	80	81	82	83	84	85
END MO	JAN ☐	FEB ☐	MAR ☐	APR ☐	MAY ☐	JUN ☐	JUL ☐	AUG ☐	SEP ☐	OCT ☐	NOV ☐	DEC ☐
YR	74	75	76	77	78	79	80	81	82	83	84	85

I. NONRATED PERIOD
NO. OF MONTHS ________
REASON CODES ________

J. DUTY POSITION TITLE
AUTH PAY GR

PART II RATINGS

A. BRIEF DESCRIPTION OF DUTIES

B. INDORSER HAS NOT OBSERVED AND CANNOT RATE SOLDIER ☐

C. REPORT BASED ON:

	DAILY CONTACT	FREQ OBSN	INFREQ OBSN	REPT & REC
R	☐	☐	☐	☐
I	☐	☐	☐	☐

D. SOLDIER SUPPORTS THE ARMY'S EQUAL OPPORTUNITY PROGRAM

	YES	NO
R	☐	☐
I	☐	☐

E. DUTY PERFORMANCE TRAITS

DUTY PERFORMANCE TRAITS		RANKS WITH VERY BEST	SUPERIOR TO MOST	EXCEEDS OR MEETS DUTY REQUIREMENTS		NEEDS IMPROVEMENT Some	NEEDS IMPROVEMENT Much	SCORE R	SCORE I
		5	4	3	2	1	0	R	I
1. Is well informed on all phases of assigned duties. (Scope of knowledge about duties)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
2. Carries out orders without constant supervision. (Dependability in performing without supervision)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
3. Shows interest and enthusiasm for duties. (Attitude toward duties)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
4. Demonstrates qualities of leadership. (Exerts positive influence on others)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
5. Seeks out opportunities for self-improvement. (Effort directed toward realization of potential)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
6. Displays ability to initiate action without direction from others. (Aggressive pursuit of methods to improve duty performance)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
7. Is successful in working with others. (Ability to work in harmony with others)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
8. Personal behavior sets a good example for others. (High standards of personal conduct)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
9. Takes pride in dress and appearance. (Neat and military in bearing)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
10. Is physically fit, as required, for MOS/grade during combat. (Physical condition)	R I	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐		
							TOTALS		

F. DEMONSTRATED OVERALL PERFORMANCE OF ASSIGNED DUTIES

	Ranks With Very Best		Superior to Most			Exceeds or Meets Duty Requirements				Demonstrates Shortcomings Minor			Demonstrates Shortcomings Major			SCORE
R	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
	44	43	42	38	34	33	27	21	15	14	10	6	5	3	1	
I	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

G. ADVANCEMENT POTENTIAL

IF I HAD THE AUTHORITY AND RESPONSIBILITY TO DO SO, I WOULD: (DISREGARD TIME IN GRADE REQUIREMENTS)

	Promote Immediately		Promote Ahead of Peers			Promote With Peers				Not Promote			Deny Continued Active Duty	SCORE
R	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
	31	30	28	26	24	22	18	14	10	7	5	3	0	
I	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

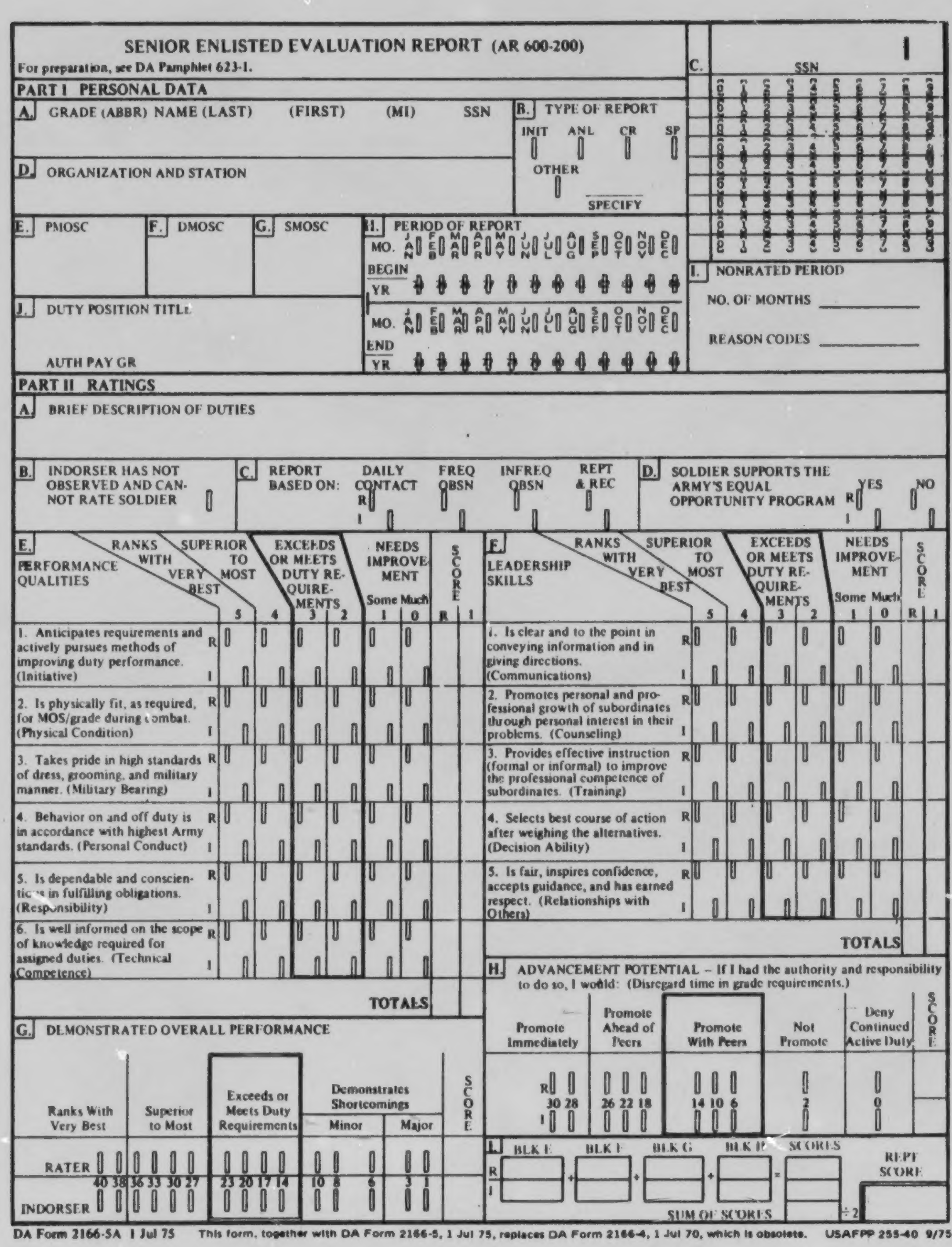
H. SCORE

BLOCKS	RATER		INDORSER				REPT SCORE
E	☐		☐				
F	☐		☐				
G	☐		☐				
SUM	☐	+	☐	=	☐	÷ 2 =	☐

DA Form 2166-5 1 Jul 75 This form, together with DA Form 2166-5A, 1 Jul 75, replaces DA Form 2166-4, 1 Jul 70, which is obsolete.

"FOR INSTRUCTIONAL PURPOSES ONLY"

Figure 5



SENIOR ENLISTED EVALUATION REPORT (AR 600-200)

For preparation, see DA Pamphlet 623-1.

PART I PERSONAL DATA

A. GRADE (ABBR) NAME (LAST) (FIRST) (MI) SSN

B. TYPE OF REPORT

INIT ☐ ANL ☐ CR ☐ SP ☐

OTHER ☐ SPECIFY

C. SSN

0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9

D. ORGANIZATION AND STATION

E. PMOSC

F. DMOSC

G. SMOSC

H. PERIOD OF REPORT

BEGIN MO. JAN ☐ FEB ☐ MAR ☐ APR ☐ MAY ☐ JUN ☐ JUL ☐ AUG ☐ SEP ☐ OCT ☐ NOV ☐ DEC ☐

YR 0 1 2 3 4 5 6 7 8 9 0 1

END MO. JAN ☐ FEB ☐ MAR ☐ APR ☐ MAY ☐ JUN ☐ JUL ☐ AUG ☐ SEP ☐ OCT ☐ NOV ☐ DEC ☐

YR 0 1 2 3 4 5 6 7 8 9 0 1

I. NONRATED PERIOD

NO. OF MONTHS ____________

REASON CODES ____________

J. DUTY POSITION TITLE

AUTH PAY GR

PART II RATINGS

A. BRIEF DESCRIPTION OF DUTIES

B. INDORSER HAS NOT OBSERVED AND CANNOT RATE SOLDIER ☐

C. REPORT BASED ON: DAILY CONTACT R ☐ I ☐ FREQ OBSN ☐ ☐ INFREQ OBSN ☐ ☐ REPT & REC ☐ ☐

D. SOLDIER SUPPORTS THE ARMY'S EQUAL OPPORTUNITY PROGRAM YES R ☐ I ☐ NO ☐ ☐

E. PERFORMANCE QUALITIES

PERFORMANCE QUALITIES		RANKS WITH VERY BEST 5	SUPERIOR TO MOST 4	EXCEEDS OR MEETS DUTY REQUIREMENTS 3	2	NEEDS IMPROVEMENT Some 1	Much 0	SCORE R	I
1. Anticipates requirements and actively pursues methods of improving duty performance. (Initiative)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
2. Is physically fit, as required, for MOS/grade during combat. (Physical Condition)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
3. Takes pride in high standards of dress, grooming, and military manner. (Military Bearing)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
4. Behavior on and off duty is in accordance with highest Army standards. (Personal Conduct)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
5. Is dependable and conscientious in fulfilling obligations. (Responsibility)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
6. Is well informed on the scope of knowledge required for assigned duties. (Technical Competence)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
							TOTALS		

F. LEADERSHIP SKILLS

LEADERSHIP SKILLS		RANKS WITH VERY BEST 5	SUPERIOR TO MOST 4	EXCEEDS OR MEETS DUTY REQUIREMENTS 3	2	NEEDS IMPROVEMENT Some 1	Much 0	SCORE R	I
1. Is clear and to the point in conveying information and in giving directions. (Communications)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
2. Promotes personal and professional growth of subordinates through personal interest in their problems. (Counseling)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
3. Provides effective instruction (formal or informal) to improve the professional competence of subordinates. (Training)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
4. Selects best course of action after weighing the alternatives. (Decision Ability)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
5. Is fair, inspires confidence, accepts guidance, and has earned respect. (Relationships with Others)	R	☐	☐	☐	☐	☐	☐		
	I	☐	☐	☐	☐	☐	☐		
							TOTALS		

G. DEMONSTRATED OVERALL PERFORMANCE

	Ranks With Very Best		Superior to Most				Exceeds or Meets Duty Requirements				Demonstrates Shortcomings Minor			Major		SCORE
RATER	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
	40	38	36	33	30	27	23	20	17	14	10	8	6	3	1	
INDORSER	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

H. ADVANCEMENT POTENTIAL – If I had the authority and responsibility to do so, I would: (Disregard time in grade requirements.)

	Promote Immediately		Promote Ahead of Peers			Promote With Peers			Not Promote	Deny Continued Active Duty	SCORE
R	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	
	30	28	26	22	18	14	10	6	2	0	
I	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	

I.

	BLK E		BLK F		BLK G		BLK H		SCORES		REPT SCORE
R		+		+		+		=			
I		+		+		+		=			
									SUM OF SCORES	÷ 2	

DA Form 2166-5A 1 Jul 75 This form, together with DA Form 2166-5, 1 Jul 75, replaces DA Form 2166-4, 1 Jul 70, which is obsolete. USAFPP 255-40 9/75

Figure 6

RATED SOLDIER'S LAST NAME AND SSN

PART II CONTINUED

I. CAREER DEVELOPMENT (RECOMMENDATIONS ON SCHOOLING AND ASSIGNMENTS)

J. 1. COMMENTS ARE MANDATORY TO JUSTIFY RATINGS IN PART II AS FOLLOWS:

a. BLOCK E SCORE BELOW 10 OR OVER 40, BLOCK F SCORE BELOW 6 OR OVER 42, BLOCK G SCORE BELOW 10 OR OVER 22, OR BLOCK D IF SOLDIER DOES NOT SUPPORT ARMY'S EQUAL OPPORTUNITY PROGRAM.

b. INDORSER WHO CHECKS BLOCK II B.

2. REMARKS OTHERWISE OPTIONAL.

RATER

INDORSER

PART III RATER AUTHENTICATION

A. ORGANIZATION AND DUTY ASSIGNMENT	B. NAME AND GRADE	C. DATE
	D. SIGNATURE	

PART IV INDORSER AUTHENTICATION

A. ORGANIZATION AND DUTY ASSIGNMENT	B. NAME AND GRADE	C. DATE
	D. SIGNATURE	

PART V SOLDIER AUTHENTICATION

A. I HAVE SEEN A COPY OF THIS REPORT COMPLETE THROUGH ACTION BY THE INDORSER. I HAVE BEEN COUNSELED CONCERNING THE REPORT.	B. NAME AND GRADE	C. DATE
	D. SIGNATURE	

PART VI REVIEWER AUTHENTICATION

A. SOLDIER WAS RATED BY CORRECT RATER AND INDORSER. NO FURTHER ACTION REQUIRED. ☐

MY REVIEW RESULTS IN ACTION INDICATED BY INCLOSURES. ☐

B. ORGANIZATION AND DUTY ASSIGNMENT	C. NAME AND GRADE	D. DATE
	E. SIGNATURE	

PART VII MILPO CERTIFICATION

A. DATE REPORT ENTERED ON DA FM 2-1:	C.	D.
B. SOLDIER'S COPY GIVEN TO SOLDIER ☐ FORWARDED TO SOLDIER ☐ MAILED TO SOLDIER ☐ CERTIFIED MAIL NO. ________	MILPO SIGNATURE	UIC

1AFPP 198-24 7/78

Figure 7

RATED SOLDIER'S LAST NAME AND SSN

PART II CONTINUED

J. CAREER DEVELOPMENT (RECOMMENDATIONS ON SCHOOLING AND ASSIGNMENTS)	RECOMMEND FOR: CSM (E–8 & E–9) R ☐ I ☐ 1SG (E–7) R ☐ I ☐

K. 1. COMMENTS ARE MANDATORY TO JUSTIFY RATINGS IN PART II AS FOLLOWS:
a. BLOCK E SCORE BELOW 6 OR OVER 24, BLOCK F SCORE BELOW 5 OR OVER 20, BLOCK G SCORE BELOW 6 OR OVER 36, BLOCK H SCORE BELOW 6 OR OVER 14, OR BLOCK D IF SOLDIER DOES NOT SUPPORT ARMY'S EQUAL OPPORTUNITY PROGRAM.
b. INDORSER WHO CHECKS BLOCK II B.
2. REMARKS OTHERWISE OPTIONAL.

RATER

INDORSER

PART III RATER AUTHENTICATION

A. ORGANIZATION AND DUTY ASSIGNMENT	**B.** NAME AND GRADE	**C.** DATE
	D. SIGNATURE	

PART IV INDORSER AUTHENTICATION

A. ORGANIZATION AND DUTY ASSIGNMENT	**B.** NAME AND GRADE	**C.** DATE
	D. SIGNATURE	

PART V SOLDIER AUTHENTICATION

A. I HAVE SEEN A COPY OF THIS REPORT COMPLETE THROUGH ACTION BY THE INDORSER. I HAVE BEEN COUNSELED CONCERNING THE REPORT.	**B.** NAME AND GRADE	**C.** DATE
	D. SIGNATURE	

PART VI REVIEWER AUTHENTICATION

A. SOLDIER WAS RATED BY CORRECT RATER AND INDORSER. NO FURTHER ACTION REQUIRED. ☐
MY REVIEW RESULTS IN ACTION INDICATED BY INCLOSURES. ☐

B. ORGANIZATION AND DUTY ASSIGNMENT	**C.** NAME AND GRADE	**D.** DATE
	E. SIGNATURE	

PART VII MILPO CERTIFICATION

A. DATE REPORT ENTERED ON DA FM 2-1:	**C.**	**D.**
B. SOLDIER'S COPY: GIVEN TO SOLDIER ☐ FORWARDED TO SOLDIER ☐ MAILED TO SOLDIER ☐ CERTIFIED MAIL NO. ____________	MILPO SIGNATURE	UIC

The Evaluation Reports normally would be completed and processed in accordance with procedures and responsibilities shown in this flow chart, Figure 8. The form is initiated by the Military Personnel Officer and is completed, in turn, by the rater and indorser. The indorser is responsible for obtaining the soldier's signature and for forwarding the Evaluation Report to the reviewer who, in turn, returns it to the MILPO for processing. As noted earlier, the forms have been designed in coordination with the SIDPERS people in PERSINSD so that all of Part I can be filled in by computer. This should decrease MILPO work and drastically reduce the error rate which we now have, most of which relates to the information in Part I. We now have at any one time in EEC between 20,000 and 30,000 reports which cannot be processed and have to be returned to the field for correction. As a result, the latest data is sometimes not available on soldiers who are pending some personnel action. The flow for adverse reports is modified to the extent that the report is forwarded to the reviewer prior to being signed by the soldier. The reviewer has generally the same responsibilities currently applicable for OERs.

We have prepared an instructional pamphlet for the rater, indorser, and reviewer. An AR will, of course, be published but we believe that the NCOs and junior officers who administer the system frequently will find that this is not an available or suitable reference. To assist in the preparation of the Enlisted Evaluation Reports, we propose that a pamphlet which gives brief, direct, and readable instructions be made available to all raters, indorsers, and reviewers.

The two forms have been extensively field tested and have been subjected to a number of critiques. The field trial was conducted at eight major CONUS installations, two Headquarters, and several Reserve Component Units.

The sample was selected so as to be representative by pay grade and included over 3,000 soldiers.

In addition to the actual EERs and SEERs, we received completed questionnaires and comments from most of the field test participants. The general reaction to the proposed form was overwhelmingly favorable. Eighty-eight percent of the raters and eighty-nine percent of the indorsers regarded the form as a definite improvement over the old form. There were, of course, numerous comments concerning specific aspects of the form. These recommendations have resulted in substantial modifications which have been incorporated in our present recommended version. The following Figure 9 shows the actual numerical results achieved by the field trial.

The average scores achieved by all pay grades in the field trial were substantially lower than the average scores on the present EER.

Figure 8

FLOWCHART FOR ADMINISTRATIVE PROCESSING

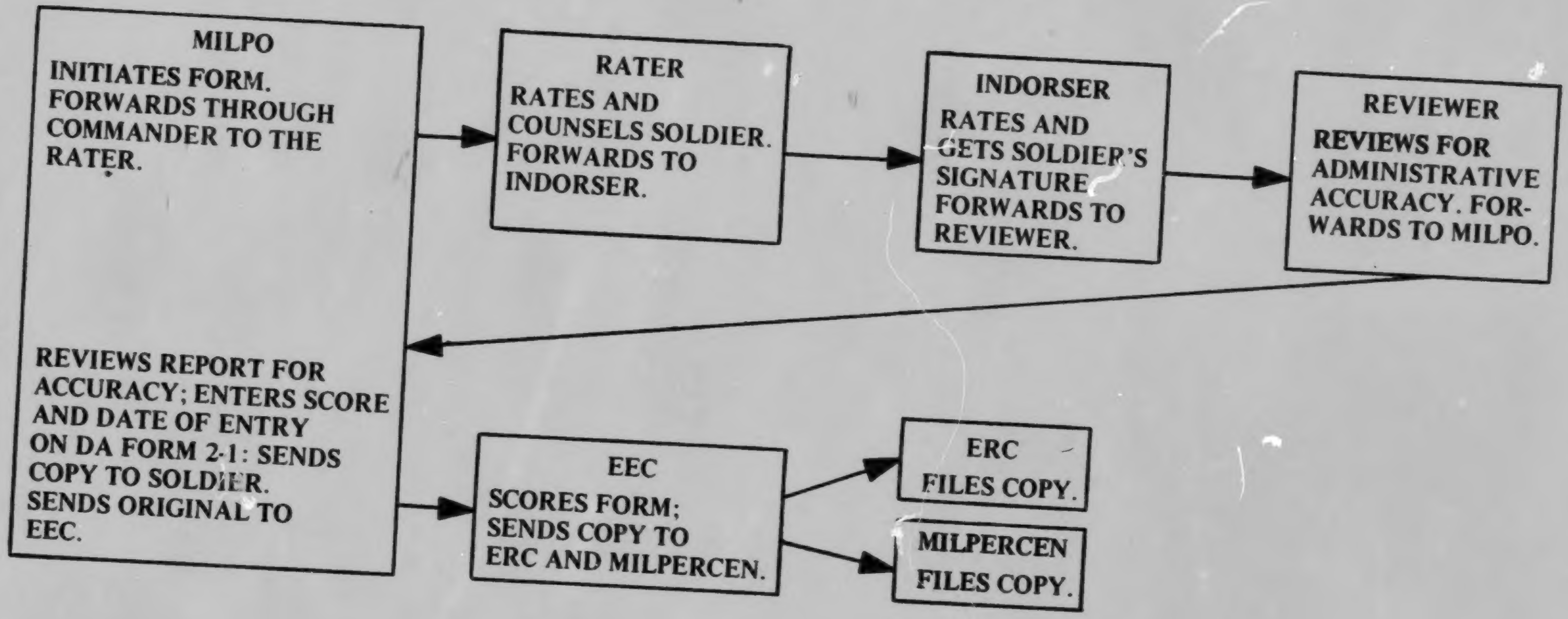

MILPO
INITIATES FORM. FORWARDS THROUGH COMMANDER TO THE RATER.
REVIEWS REPORT FOR ACCURACY; ENTERS SCORE AND DATE OF ENTRY ON DA FORM 2-1: SENDS COPY TO SOLDIER. SENDS ORIGINAL TO EEC.
RATER
RATES AND COUNSELS SOLDIER. FORWARDS TO INDORSER.
INDORSER
RATES AND GETS SOLDIER'S SIGNATURE FORWARDS TO REVIEWER.
REVIEWER
REVIEWS FOR ADMINISTRATIVE ACCURACY. FORWARDS TO MILPO.
EEC
SCORES FORM; SENDS COPY TO ERC AND MILPERCEN.
ERC
FILES COPY.
MILPERCEN
FILES COPY.

Figure 9

COMPARISON OF 1974 EER AND FIELD TRIAL EER/SEER

PAY GRADE	CURRENT EER ARMY AVERAGE SCORE	EER/SEER AVERAGE SCORE
E3	87.2	76.9
E4	106.5	92.1
E5	110.2	96.9
E6	116.9	102.4 EER 98.4 SEER
E7	120.6	106.0
E8	122.7	110.8
E9	123.7	112.3

We believe that at least for a while the new forms will dampen inflationary trends and improve our ability to discriminate among soldiers. Several factors contribute to this: First, just the use of a new form combined with a good education program prior to implementation should help. Next, we believe that there will be a psychological impact resulting from the use of two forms where the raters compare senior NCOs with respect to other senior NCOs rather than with respect to the Army as a whole. Finally, the structure of the form facilitates finer tuning of the ratings at the upper end of the spectrum than does the current form.

One point which was made emphatic to us by the field test results was the unacceptability of a multicopy evaluation form. A multicopy form was initially planned to eliminate the Xeroxing requirements and to have useful copies for all required files, as well as for the soldier. But, the multiform approach was almost universally rejected. Some of the more frequent comments included difficulty in aligning the form, that it tears easily, is bulky, and is costly. In addition, Enlisted Records Center indicated that the carbon copy that they were supposed to receive was unacceptable for permanent records.

In addition to the field test, the new EERs have been critiqued by a class at the SGM Academy, the Enlisted Records Center, E8 Selection Board, as well as the various groups who had been briefed in connection with the approval process. These group comments have been incorporated into the forms.

The forms were also briefed earlier to the SGM's Conference in Washington. Support from this group was enthusiastic. They particularly liked the forms as an improved tool for counselling as well as for the requirement for rater counselling and soldier acknowledgement built into the system.

In conclusion, we are enthusiastic that these new Enlisted Evaluation Reports will be effective. We believe that there are significant improvements over the present Enlisted Efficiency Report. In October of this year the new EER and SEER will be in use throughout the US Army. Evaluation of their effectiveness depends upon their proper utilization and the eternal hope Behavioral Scientists hold for making something work.

Early Identification of Maladaptive Accessions

Jeffrey E. Kantor and Nancy Guinn
Personnel Research Division
Air Force Human Resources Laboratory

Identification of personnel who are most likely to succeed on the job is the primary goal of every selection program. Hiring personnel who will later terminate employment due to nonadaptability represents a cost which might be minimized if more precise selection procedures could be developed. All branches of the Armed Services have been concerned with problems in adaptation, and for over 15 years research has been directed at these problems (Flyer 1959; Klinger, Dubiusson, and deJung, 1961; Plag and Goffman, 1966).

In 1972, Air Force medical personnel initiated a research project to develop a screening technique which could be used to identify recruits who are considered high-risk for problems in adaptation to the military environment (LaChar, Larson, and Sparks, 1974). For use in this project, LaChar, et. al., developed a 100-item self-report History Opinion Inventory (HOI) designed to tap dimensions of school adjustment, family stability, social orientation, emotional stability, bodily complaints, motivation and expectations for achievement, and response toward authority.

Using the inventory and criterion data obtained on approximately 15,000 male airmen during basic training, two predictive scales were developed from the HOI for future use in screening. The PEI scale was designed to measure characteristics associated with emotional stability; the PDA scale was designed to measure those characteristics associated with admission of previous drug usage.

These two scales were then combined into an Adaptation Index (ADI), and an optimal cut-off score was determined which would classify recruits into one of two categories, normal or high-risk. Based on this ADI cut-off score, 12 percent of the sample population was labeled as high-risk for military adaptation. According to records maintained by LaChar, et. al., half of the high-risk group did, in fact, experience problems in adjustment during basic training although their problems did not necessarily result in discharge from service. This high-risk group was composed of seven percent of the normal criterion group, 41 percent of the severe adjustment group, and 47 percent of the drug discharge group.

Based on these results, it was concluded that prediction of initial adaptability to military service is possible, and that such screening could result in substantial savings to the Air Force in identifying personnel who require special treatment or who should be separated from service (LaChar, et. al., 1974). However, the criterion classifications used in the initial analyses were partially based on subjective clinical and instructor evaluations of the individual's behavior. Prior to consideration of such an instrument for use in the operational screening of Air Force accessions, it was considered advisable that further investigation of the History Opinion Inventory be accomplished using the objective criterion of in-service versus actual separation/discharge from service.

The objectives of the current study were to follow-up the accessions administered the History Opinion Inventory in basic military training in order to determine the accuracy of the HOI scores in predicting the criterion of in/out of service during the first two years after enlistment and to determine whether additional aptitudinal and biographical data might increase the effectiveness of the screening procedure.

Method

Subjects: The sample population consisted of 15,252 basic airmen who were administered the HOI during basic military training at Lackland AFB from June through August 1972.

Procedures: The data files established by LaChar (1974) were matched with the airman tape files maintained by the Computational Sciences Division of the Air Force Human Resources Laboratory to obtain aptitudinal, biographical, and discharge data. A total of 160 cases in the original population did not match the official data files which reduced the sample population to 15,092.

Discharge status was determined by a standard designation number (i.e., loss code) which identified all personnel who had been separated or discharged from service during the first two years after enlistment.

On the basis of discharge status, two groups of subjects were formed. The In-service group was composed of 10,329 airmen who were still on active duty or had extended their original commitment as of September 1974. The Out-of-service group was composed of 4,764 airmen who had been separated or discharged from service during their initial two years.

Scoring of the HOI response data and the cut-off scores used in the current study are those previously established by LaChar, et. al., (1974). The actual items, scoring used in deriving the HOI scale scores and estimates of scale reliability, are presented in Tables A1 and A2 in the appendix. The weighted linear combination of the two scales used to derive the Adaptation Index was .6568 of the PEI scale value and .7541 of the PDA scale value. The decision rules (cut-off scores) based on the optimum value which differentiated between recruits who would and would not have problems in basic training were as follows: PEI - 7.5 scale value; PDA - 11.5 scale value; ADI - 12.5 scale value.

Tests of significance between group means were accomplished utilizing t-tests. Correlation procedures were used to indicate the relationship between HOI scores and the criterion, and regression analyses were accomplished to determine the usefulness of biographical and aptitudinal data in predicting adaptability to military service and whether these data significantly increase accuracy in prediction over and above the use of the HOI scores alone.

Results and Discussion

The means and standard deviations of the HOI scores, by group, are presented in Table 1. Results of the t-tests between the group means are also presented in Table 1. For each scale, the out-of-service group scale score mean was significantly higher than the in-service group scale score means.

Correlation coefficients between the HOI scores and the criterion were: PEI - .2022; PDA - .2735; and ADI - .2659. All correlations were significant beyond the .01 level.

Statistically then, the HOI appears to hold some promise of functioning in a useful manner as a screening device. However, since statistical significance can fail to reflect the practical usefulness of a screening device, it is also important to consider hit and false positive rates generated by a screening paradigm. Therefore, Table 2 shows the percentage of each group identified as "normal" or "high risk" using the decision rules established by LaChar in 1974. On the PEI scale only 6% of the in-service group were incorrectly labeled as high risk while 18% of the out-of-service group were correctly labeled as high risk. On the PDA scale, only 5% of the in-service group were incorrectly labeled as high risk while 22% of the out-of-service group were correctly labeled as high risk. Finally, on the ADI scale, only 6% of the in-service group were incorrectly labeled while 23% of the out-of-service group were correctly labeled.

It would appear as if the PDA scale alone may be capable of functioning as effectively as the more complexly derived ADI scale. This possibility and that of increasing the accuracy of prediction by utilizing normally available aptitude and biographic data in

TABLE 1

Means, Standard Deviations, and "t" Tests of Scale Scores

		PEI Scale			PDA Scale			ADI Scale		
	Valid N	Mean	SD	"t"	Mean	SD	"t"	Mean	SD	"t"
In-service	10,329	3.03	2.42		5.20	3.46		5.91	3.80	
				23.00*			31.35*			30.43*
Out-of-service	4,764	4.16	3.51		7.46	5.25		8.35	[illegible].93	

*$\rho < .01$

TABLE 2

Percent of "Normal" and "High Risk" by Group

	PEI Scale		PDA Scale		ADI Index	
	Normal	High Risk	Normal	High Risk	Normal	High Risk
In-service	94	6	95	5	94	6
Out-of-service	82	18	78	22	77	23

conjunction with the HOI scores were investigated through regression analyses.

Multiple correlations for the various groups of predictors are presented in Table 3, and summaries of the regression analyses are presented in Table 4. The first regression problem indicates that the aptitude and biographic data do add significantly to prediction over and above the ADI score while the second regression problem indicates that the aptitude and biographic information alone do not predict as well as when used in conjunction with the ADI score. The third problem indicates that the PDA scale contributes as much as both the PEI and ADI scales together and therefore, could be used singly, negating the need to compute the derived ADI score. The final problem indicates that the aptitude and biographic information do make a significant contribution even when used with the PDA scale alone.

Conclusions and Recommendations

The correlations obtained for the HOI scales appear to indicate that the self-report data contained in the HOI has some practical usefulness as a rough, preliminary screening device. However, the small magnitude of the observed relationships necessitates careful review by professionals of all personnel identified as high-risk. In no instance should identification as high-risk by the HOI be used as the sole basis for any adverse personnel action. On the positive side, the HOI does identify a sizeable proportion of recruits who were actually discharged from service as undesirable during the first two years of active duty. However, the overall savings which might be accrued from early identification of the high-risk group might be obscured by the costs of implementing a secondary assessment and counseling phase which is considered necessary with the use of any rough screening device such as the HOI.

Prior to the use of the HOI in an operational setting, the following recommended courses of action are suggested.

a. Revalidate the HOI on current accessions to determine its effectiveness and stability on a new population. In the original sample, a large number of personnel who were discharged for admission to prior drug usage were used for scale construction. Although the PDA scale appears to be quite effective in identifying all types of undesirable losses, the appropriateness of the original scales or cut-off scores developed on that population may be questionable if a decrease in the number of drug discharges has occurred during the past two years.

b. While results obtained on the original sample suggest that such a screening procedure might be used effectively, the population

TABLE 3

Multiple Correlations

	Aptitudinal & Biographical data	Apt. & Bio & ADI	Apt., Bio, PDA, PEI	Apt., Bio & PDA
In-service/Out-of-service	.2149	.3127	.3158	.3150

TABLE 4

SUMMARY OF REGRESSION ANALYSES

	R^2-Full Model	Predictors in Full Model	R^2-Restricted Model	Predictors in Restricted Model	df_1	df_2	F Ratio
In-service/ Out-of-service	.0978	M, A, G, E, AFQT, Age, Yrs of Educ, ADI	.0707	ADI	7	15083	61.86**
	.0978	M, A, G, E, AFQT, Age, Yrs of Educ, ADI	.0462	M, A, G, E, AFQT, Age, Yrs of Educ	1	15083	824.55**
	.0750	ADI, PEI, PDA	.0748	PDA	2	15083	1.07
	.0993	M, A, G, E, AFQT, Age, Yrs of Educ, PDA	.0748	PDA	7	15083	55.88**

consisted of male accessions only. Prior to using the HOI as a screening device on a female population, additional research must be accomplished to establish the applicability of the scales and cutting scores for a female population.

c. It is further recommended that use of this screening device should be limited to preliminary screening only and that additional psychometric and/or psychiatric assessment be mandatory before any personnel action is recommended. Every effort should be made to retain as many of the potentially successful personnel in the high-risk category as possible.

d. To simplify the administrative scoring of the HOI, it is recommended that further research on this instrument investigate the possibility of developing one scale rather than the complex weighting process used for deriving the ADI index score.

e. Based on the preliminary regression analyses, additional aptitudinal and biographical data which is available on all recruits should be considered in combination with HOI data to improve the accuracy in identification of maladaptive personnel.

References

Flyer, E. S. Factors Relating to Discharge for Unsuitability Among 1956 Airmen Accessions to the Air Force. WADC-TN-59-201. Lackland AFB, Texas: Personnel Laboratory, Wright Air Development Center, December 1959.

Klieger, W. A., Dubuisson, A. U., and deJung, J. E. Prediction of Unacceptable Performance in the Army. Tech Research Note 113. Human Factors Research Branch, TAG Research and Development Command, June 1961.

LaChar, D., Sparks, J. C., and Larsen, R. M. Psychometric Prediction of Behavioral Criteria of Adaption for USAF Basic Trainees. Journal of Community Psychology, 2 (3), 1974, 268-277.

Plag, J. A., Arthur, R. J., and Goffman, J. M. Dimension of Psychiatric Illness Among First-term Enlistees in the United States Navy. Military Medicine, 1970, 135, 665-673.

Table A1. PEI SCALE - EMOTIONAL STABILITY

		Response scored as +1
1.	I have needed help for emotional problems.	T
2.	At one time I needed medication to stay calm.	T
3.	For a long time I have had difficulty sleeping.	T
4.	I often have headaches.	T
5.	I have cried several times this past year.	T
6.	I usually take things hard.	T
7.	I enjoyed physical education.	F
8.	I have had more than my share of illness.	T
9.	I needed special help with my school studies.	T
10.	I am joining the Air Force to get a better education.	T
11.	I would rather work by myself than with others.	T
12.	I was a slow learner in school.	T
13.	I would rather read than be with people.	T
14.	I entered the service (AF) because there was nothing else to do.	T
15.	I do not mind orders and being told what to do.	F
16.	As a child I was a loner.	T
17.	My father is (was) a nervous man.	T
18.	I never cared much for school.	T

KR-20 scale reliability: .716

Table A2. PDA SCALE - Drug Use Admission

		Response scored as +1
1.	I often played hookey from school.	T
2.	I quit school because I lost interest.	T
3.	I feel better when I drink.	T
4.	For a long time I have had difficulty sleeping.	T
5.	I think I will make the Air Force a career.	F
6.	I am joining the Air Force to get a better education.	F
7.	I never cared much for school.	T
8.	I have been in trouble with the police.	T
9.	I was suspended from school more than two times.	T
10.	I have often gone against my parents' wishes.	T
11.	I do not mind orders and being told what to do.	F
12.	I often have headaches.	T
13.	I entered the service (AF) because there was nothing else to do.	T
14.	I had my share of trouble with teachers.	T
15.	I was expelled or suspended from school.	T
16.	At one time I needed medication to stay calm.	T
17.	I have never done any heavy drinking.	F
18.	I enjoyed physical education.	F
19.	I quit school because I was failing.	T
20.	I have been expelled from school more than once.	T
21.	I have been arrested more than twice.	T

Table A2 (Cont'd)

	Response scored as +1
22. High school was boring.	T
23. I often cuss and swear.	T
24. I plan to attend college.	F
25. I sometimes wanted to run away from home.	T
26. I have needed help for emotional problems.	T

KR-20 scale reliability: .803

Comparative Performance of High School Graduates/Non-Graduates in Military Service

Jeffrey E. Kantor and Nancy Guinn
Personnel Research Division
Air Force Human Resources Laboratory

Previous research has shown that high school graduation, or years of education completed, has been an important correlate of success in a military career. As early as 1955, the utility of years of education as a predictive variable for initial selection was a topic of research (Humphreys, 1955). Vitola, et. al. (1967) found a positive relationship between aptitude index scores and educational status with high school non-graduates achieving lower average scores than their counterparts with more education. The usefulness of educational data in classification and assignment has also received attention (Brokaw, 1962; 1963; Lecznar, 1964) with the general finding that high school graduation was related to success in technical training. A high positive relationship between educational level and service adjustment has been documented quite extensively (Flyer, 1959, 1963; Fisher, et. al., 1960; Gorden and Bottenberg, 1962; Plag, 1962; Gunderson, 1963; Plag, et. al., 1970; Drucker and Schwartz, 1973; Shoemaker, et. al., 1974). Finally, Plag and Goffman (1966, 1967) have found a strong positive relationship between educational level and successful completion of the initial tour.

Results of studies such as these, coupled with the current, very favorable selection ratio, have provided the justification for the implementation of more restrictive enlistment policies for the Air Force. Beginning in January, 1975, a prospective recruit had to meet the following prerequisites: a total of 170 on the four combined aptitude test composites of the Armed Forces Vocational Aptitude Battery (ASVAB), a minimum score of 45 on the General Aptitude Index, and, if classified in mental Categories III or IV on the Armed Forces Qualifying Test (AFQT), completion of high school. With the overall economic situation and present unemployment rates, the Air Force has had no difficulty in attracting the numbers of high quality men and women needed to fulfill its manning requirements. However, if the selection ratio becomes less favorable sometime in the future, enlistment standards may have to be modified again to maintain a viable force level. Attention then will have to be directed towards making a policy decision which will, at once, both broaden the manpower resource base and also maximize the selection of potentially successful recruits and minimize the acceptance of those not likely to become successful servicemen.

A Navy project on administrative and disciplinary discharges found a higher disciplinary loss rate among the mental Category I high school non-graduates than among the non-graduates in the lower mental categories (Stephan, et. al., 1972). This finding indicates a need for a closer comparative evaluation of high school graduates and non-graduates in order to delineate the advantages and disadvantages of recruiting individuals with lower educational qualifications. This type of information, on performance and career progression of individuals with lower educational qualifications, should prove invaluable in planning for revisions in enlistment standards needed in times of less favorable selection ratios. This study investigates and compares the performance and career progression of a sample of high school graduates and non-graduates during their initial four-year tour of military service.

Method

A total of 20,705 male non-prior service enlisted accessions designated as Project 100,000 personnel, who enlisted during the period April 1967 through March 1968, comprised the sample population. Subjects included all accessions classified by the AFQT as Category I and IV personnel and a 10% random sample of all Category II and III personnel who enlisted during the specified time period. All data regarding the progress, performance, and retention of these airmen was obtained from the historical record files maintained by the Computational Sciences Division of the Air Force Human Resources Laboratory.

The sample population was divided by educational levels for comparative analyses. Those individuals who completed 12 years or more of education prior to enlistment comprised the high school graduate group. Those accessions indicating they had completed only 11 years or less of education were designated as high school non-graduates. Since ability level might be related to differential performance, the sample was further subdivided, by mental categories, as established by performance on the AFQT. Table 1 shows the number of individuals in each of the educational/mental category subgroups.

Subgroups of interest were compared on various criterion measures of success in the Air Force. These measures were: graduation/elimination from basic military training, graduation/elimination from technical training, number of disciplinary actions, attainment of skill level, number of unsuitability discharges, and reenlistment decision.

For each of the variables, chi-square analyses were used to test for statistical significance of any differences, and where appropriate, an overall chi-square was partitioned into its one degree of freedom components utilizing Castellian's method (Castellian, 1965). The Type I error rate was controlled at .001 per comparison.

Table 1

Sample Characteristics

Mental Category	AFQT Percentile Range	(N) High School Graduates	(N) High School Non-Graduates
I	93-99	6150	45
II	65-92	3183	69
III	31-64	3016	152
IV	21-30	7770	320

Table 2

Percentages of Educational Groups at Criteria

	BMT Successes	Tech Training Successes	No Disciplinary Actions	Unsuitability Discharges	Reenlisted After Tour
High School Graduates	96.69	94.58	95.70	8.45	23.92
High School Non-Graduates	88.23	89.54	90.10	27.35	29.86

Table 3

Percentages at Skill Levels

	Year 1 (% at Skill Level 1)	Year 2 (% at Skill Levels 1 or 3)	Year 3 (% at Skill Levels 1 or 3)	Year 4 (% at Skill Levels 1, 3, or 5)
High School Graduates	19.60	17.33	6.15	97.99
High School Non-Graduates	25.35	35.45	15.35	98.74

Results and Discussion

One of the first measures of success in military service is the successful completion of the six-week basic military training (BMT) program. Comparing the successful completion rates between high school graduates and non-graduates, it was found that a significantly higher percentage of high school graduates successfully completed BMT than did the non-graduates. These percentages are presented in Table 2. Additionally, among the high school non-graduates, it was found that significantly fewer (83.13%) non-graduate mental Category IV airmen completed BMT than did non-graduate mental Category III airmen (94.08%).

After BMT, approximately 80% of Air Force accessions are assigned to some form of technical training. This represents a considerable cost in general, and, in particular, the funds expended on individuals who do not successfully graduate from their technical training programs is money wasted. Therefore, any differential performance associated with educational level becomes a matter of concern. Comparing the successful completion rates for technical training, it was found, again, that significantly higher percentages of high school graduates completed technical training than did the non-graduates. These percentages are also presented in Table 2. Among the non-graduates, there were no significant differences associated with mental category on technical training graduation.

Another way to assess differential performance between groups in the military is by monitoring their progression through successively higher skill levels. The Air Force has a formalized series of skill levels which describe the working capability of an airman. These are comparable across differing jobs, and each succeeding level encompasses both greater specific ability and general responsibility. For this study, the skill levels of interest were: Skill Level 1, "unskilled;" Skill Level 3, "apprentice;" Skill Level 5, "Journeyman;" and Skill Level 7, "supervisor." The first skill level comparison was made after the completion of the initial year of service, and succeeding comparisons were made at one-year intervals with the last comparison involving the skill level obtained or held at the completion of the airman's initial four years of service. The skill level comparisons were made on the following dichotomies: for the first year, Skill Level 1 versus all advanced skill levels; for the second and third years, Skill Level 1 and 3 versus 5 and 7; and for the fourth year, Skill Levels 1, 3, and 5 versus 7. These breakpoints were chosen because they appeared to represent, respectively, atypical lag, normal progression, and atypical advancement.

The percentages of high school graduate and non-graduate airmen still at the lower skill levels for each comparison are shown in

Table 3. For the first three comparisons, between graduates and non-graduates, significantly larger percentages of high school non-graduates were still at the lower skill levels. There is no significant difference for the fourth comparison although the difference is in the same direction as in the previous comparisons. Among the high school non-graduates, only in the second year was any significant difference found associated with mental categories. In this instance, a higher percentage of mental Category IV and III non-graduate airmen were still at lower skill levels than mental Category II non-graduate airmen. Aside from that specific case, no differences among high school non-graduate airmen were found.

Anytime an airman fails to complete his obligated tour, he represents a dollar and resource loss which might have been prevented. Disciplinary actions often lead to premature separation and are, in themselves, an additional burden in terms of money and time. Therefore, the percentages of high school graduate and non-graduate airmen completing their tour without being involved in any disciplinary action is a matter of concern. These percentages are also shown in Table 2. It was found that a significantly higher percentage of high school graduates completed their tour without any disciplinary action than did high school non-graduates. No differences were found among the non-graduate airmen.

A direct indication of premature loss is an unsuitability discharge. The percentages of high school graduate and non-graduate airmen receiving a premature discharge because of unsuitability are shown in Table 2. A significantly higher percentage of non-graduate airmen received unsuitability discharges than did high school graduates. Among the non-graduate airmen, no differences were found associated with mental categories.

A final indicator of success in a military career is a favorable career decision at the end of the initial tour. The percentages of high school graduate and non-graduate airmen eligible to reenlist who did reenlist were compared and are also presented in Table 2. Significantly higher percentages of high school non-graduates reenlisted than did the high school graduates.

Summary of Results

On all measures except the last, involving reenlistment decision, the high school graduate airmen performed significantly better than did their high school non-graduate counterparts. Higher percentages of high school graduates successfully completed basic military and technical training, advanced to higher skill levels quicker, and encountered less disciplinary problems; a smaller percentage received an unsuitability discharge. It seems consistently clear, from the data, that high school graduates become more productive enlisted personnel than do high school non-graduates.

Comparisons among the high school non-graduates, by mental category, revealed only two instances where ability, as measured by the AFQT, was related to differential performance. In these instances, it was found that mental Category IV non-graduate airmen were eliminated from basic military training at a higher rate than other non-graduate airmen, and mental Category III and IV non-graduate airmen did not, at the end of their second year of service, advance as quickly to higher skill levels as did the other non-graduate airmen. Aside from these specific exceptions, no differences in performance related to mental category were found among the non-graduate airmen. However, current enlistment standards do make a distinction among those non-graduate airmen categorized as either Group I or II and those categorized as either Group III or IV. At present, only the non-graduate Group I and II applicants are eligible for enlistment. It would not seem that the data from this study would strongly support this particular cut-off point for high school non-graduate airmen since on the majority of measures there were no differences between the performance of the different mental category groups of non-graduate airmen.

Impact of Current Enlistment Standards

Current Air Force enlistment standards would not permit the enlistment of 5,537 airmen included in the sample used in this study. To allow an appraisal of the overall effectiveness of the new Air Force recruiting policy, the percentages of correct identification of unsuccessful personnel (i.e., "hit rate") and the percentages of incorrect labeling of successful personnel as potential failures (i.e., "false positive rate") are presented for several criteria in Table 4. Also, for comparative purposes, similar percentages are shown as if screening were accomplished dependent only upon the high school completion variable and from a combination of current enlistment standards plus an unequivocal high school completion requirement.

The current enlistment standards generate an impressive hit rate of approximately 40%; however, this is coupled with a false positive rate of approximately 25%. This would mean that for every three potentially unsuccessful candidates screened out from service, two potentially successful applicants are also excluded from enlisting. Only under conditions providing a very favorable selection ratio could a false positive rate this high be tolerated. In comparison, screening only on high school graduation provides the smallest hit and false positive rates while the addition of the unequivocal high school

Table 4

Impact of Differing Enlistment Standards

Enlistment Standards	Current Enlistment Standards		High School Graduation Only		Current Enlistment Standards and High School Graduation	
	Hit Rate	False Positive Rate	Hit Rate	False Positive Rate	Hit Rate	False Positive Rate
BMT	40.90	26.24	9.39	2.59	41.51	26.73
Technical Training	38 37	23.36	4.18	2.09	39.40	24.05
Criteria:						
Disciplinary Actions	41.12	26.20	6.52	2.66	41.12	26.42
Unsuitability Discharges	39.25	24.90	8.02	2.09	40.54	25.13

graduation requirement to the current enlistment standards produce a slight rise in hit rate and even smaller rise in false positive rate over those of the current enlistment standards alone.

Conclusion

The results of this study indicate significant and consistent differences in measures of military service performance between high school graduate and high school non-graduate airmen. The high school non-graduate airmen do not perform up to the standards of their high school graduate counterparts in either basic military training, technical training, or on-the-job performance as measured by attainment of advanced skill levels. Additionally, they accrue a higher percentage of disciplinary actions and unsuitability discharges. It would appear then, that the data from this study, as in previous work, reflect the importance of high school graduation for success in a military career.

However, it does not seem likely that this relationship between high school graduation and military performance can be totally explained by postulating a lower level of mental ability for the non-graduate airmen. Comparisons among the mental category sub-groups of non-graduate airmen revealed, with only two exceptions, no differences in performance related to differences in mental ability. Therefore, factors other than measured mental ability must account for the observed differences in performance. The data from this study do not allow the delineation of these other influences, but it is possible to speculate that motivational and social factors could play a large part in accounting for these differences in performance. Future research might be directed at identifying and quantifying these additional aspects as they relate to success in a military career.

The current enlistment standards, when applied a posteriori to this sample, generated impressively high hit rates but also generated high false positive rates. If conditions in the civilian sector change and effectively alter the present selection ratio, it might be possible or necessary to again change the enlistment requirements. If the selection ratio becomes more favorable, a high school graduation requirement for all mental category groups might be considered. When this restriction was applied in conjunction with current enlistment standards, a small rise in hit rate was generated (coupled with an even smaller rise in false positive rate). This additional restriction alone would not make a dramatic impact, but considering the lack of differential performance found associated with mental categories among non-graduate airmen, it would appear that this would be a logical step. Additional research would be useful in specifying what further restrictions would function in the most appropriate manner. However, if conditions change so as to produce a less fa[illegible]e selection ratio and the Air Force can no longer tolerate the high false positive rate generated by the current enlistment standards, then enlistment requirements will have to be lowered in order to fulfill manning needs. The results from this study do not lead to any clear decision concerning which high school non-graduates would be the most potentially successful sub-group of non-graduate applicants, but what few differences did surface appear to indicate that mental Category I, II, and III non-graduates would be a slightly better choice than mental Category IV non-graduates. Again, additional research is needed to delineate what other factors besides mental ability influence the course of a high school non-graduate's service career.

In summary, high school graduates constitute a more successful military group than do high school non-graduates, and among the non-graduates, in terms of mental ability sub-groups, there are almost no differences in performance.

REFERENCES

Brokaw, L. D. Prediction of technical school success from homogeneous biographical inventory scores. PRL-TDR-62-12 AD 289 877. Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, July 1962.

Brokaw, L. D. Prediction of success in technical training from self-report information on educational achievement. PRL-TDR-63-11. Lackland AFB, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, April 1963.

Castellian, W. J. On the partioning of contingency tables. Psychological Bulletin, Vol 64, (5), 1965, 330-338.

Fisher, W. E., Ward, J.H., Jr., Holdrege, F. E., and Lawrence, H. G. Prediction of unsuitability discharges. WADD-TN-60-260. Lackland AFB, Texas: Personnel Laboratory, Wright Air Development Division, October 1960.

Flyer, Eli S. Factors relating to discharge for unsuitability among 1956 airman accessions to the Air Force. Lackland AFB, Texas: Personnel Laboratory Wright Air Development Center, December 1959.

Flyer, E. S. Prediction of unsuitability among first-term airmen from aptitude indexes, high school reference data, and basic training evaluations. PRL-TDR-63-17 AD 420 530. Lackland AFB, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, June 1963.

Gordon, M. A., and Bottenberg, R. A. Prediction of unfavorable discharge by separate educational levels. PRL-TDR-62-5 AD 280 802. Lackland AFB, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, April 1962.

Gunderson, E. K. E. Biographical indicators of adaption to Naval service. Report Number 63-19. San Diego, Calif.: US Navy Medical Neuropsychiatric Research Unit, Bureau of Medicine and Surgery, November 1963.

Humphreys, L. G. Significance of years of education for airman assignments. PRL-TM-55-24. Lackland AFB, Texas: Personnel Research Laboratory, Air Force Personnel and Training Research Center, August 1955.

Lecznar, W. B. Years of education as a predictor of technical training success. PRL-TDR-64-2. Lackland AFB, Texas: 6570th Personnel Research Laboratory, Aerospace Medical Division, February 1964.

Plag, J. A. Pre-enlistment variables related to the performance and adjustment of Navy recruits. Journal of Clinical Psychology, Volume 15, 1962, 168-171.

Plag, J. A., and Goffman, J. M. The prediction of four-year military effectiveness from characteristics of Naval recruits. Military Medicine, Volume 131,2 1966, 729-735.

Plag, J. A., and Goffman, J. M. A formula for predicting effectiveness in the Navy from characteristics of high school students. Psychology in the Schools, Volume 3, 216, 222, March 1967.

Plag, J. A., Arthur, R. J., and Goffman, J. M. Dimensions of psychiatric illness among first-term enlistees in the United States Navy. Military Medicine, Volume 135, 8, 1970, 665-673.

Shoemaker, W. B., Drucker, E. H., and Kriner, R. E. Prediction of delinquency among Army enlisted men: A multivariate analysis. Technical Report 74-3. Alexandria, Virginia: Human Resources Research Organization, February 1974.

Stephan, R. A., Carroll, H. E., and Brown, N. Comparison of Navy and Air Force Administrative and Disciplinary Discharges. Center for Navy Analyses, Institute of Naval Studies. Research Contribution 218. Arlington, Virginia: Center for Naval Analyses, Institute of Naval Studies, November 1972.

Vitola, B. M., Valentine, L. D., and Tupes, E. C. Aptitude and educational data for Air Force enlistees, 1962 through 1965. PRL-TR-67-8. Lackland AFB, Texas: Personnel Research Laboratory, Aerospace Medical Division, August 1967.

AN EXPERIMENTAL PRE-ENLISTMENT SCREENING PROCEDURE*

Ted M.I. Yellen

Navy Personnel Research and Development Center
San Diego, California

During FY 74, 12.5 percent of all Navy recruits were discharged from the service before they completed basic training. Although the attrition rate has dropped below 10 percent for FY 75, this still represents a considerable number of enlistees who fail to complete basic training. There are various reasons as to why so many recruits are prematurely discharged. However, the primary cause of discharge during basic training is character and behavior disorders and inaptitude. It is apparent that current enlistment procedures are not screening out these high risk applicants.

This paper describes a research effort aimed at evaluating the Delinquent Behavior Inventory as a pre-enlistment screening tool. The research is documented in a technical report entitled "Validation of the Delinquent Behavior Inventory as a Predictor of Basic Training Attrition" (Yellen, 1975). The Delinquent Behavior Inventory was designed to identify delinquent behaviors, including illicit drug use while in the Navy. The instrument contains 102 multiple choice items which tap background, attitudes, and personality factors.

The instrument, which we call the DBI for short, was administered to 2,500 Navy recruits during their first week of basic training in San Diego. Of the 2,500 recruits, 101 were discharged from basic training for detrimental causes. For the three month period covered in this study the percentage discharged for various causes are as follows: eight percent for fraudulent entry; four percent for general unfitness; 23 percent for character and behavior disorders; and 65 percent for inaptitude.

A scoring key was developed from an item analysis program which contrasts the rate of endorsement for each item response alternative for high and low criterion groups. The basic training graduates and dropouts formed the criterion groups. The validity coefficient for the criterion of basic training graduates versus dropouts was .25. Since the "N" for the low criterion group in this study was relatively small (101), cross-validation was not performed. Typically, cross-validation coefficients are smaller than key development coefficients. In this case it means that if the DBI were used on any other sample of recruits, e.g., if used operationally the relationship may be less than .25. On the other hand the validity may improve over time since the immaturity of the data may be causing the low relationship. The criterion categorization was based on Navy Training Center graduates versus dropouts. The dropouts, which composed the low criterion group, were those individuals who were discharged for detrimental reasons. The high criterion group consisted of all the remaining recruits. As a result of this categorization, the graduate group is composed of an unknown number of Navy men who also may be found unsuitable at some time following their completion of basic training. Follow up research will be conducted in which attrition data and supervisory evaluations will be obtained on each individual. The evaluations will be concerned with disciplinary problems, if any, overall performance and general suitability. Based on these evaluations, the sample will be split into high and low criterion groups, then item validation and cross-validation analysis will be performed on the DBI data.

---

*The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Insert Table 1 about here

Table 1 portrays the distribution of DBI scores for the recruit graduate and dropout samples. If this table can be considered a stable depiction of the situation, it shows for example, that using a DBI cutting score of -1.50 would have eliminated approximately 52 percent of the dropouts (N=52) at the cost of erroneously rejecting about 13 percent of the recruit training graduates (N=300). Realistically this would be too high a loss rate. However, applying a cutting score of -18.76 would have disqualified approximately 12 percent of the basic training dropouts and the rejection of less than one percent of the graduates.

The DBI was readministered to several basic training classes of these same recruits during their fourth week of training. The reliability of the instrument was .78.

An analysis of the graduate/dropout item responses indicated a negative response pattern for the dropouts. The item alternatives selected more often by the dropouts were grouped into behavioral clusters and are shown in Table 2. They are interpreted only as indicating a trend, that is, although both groups exhibited some degree of the behavioral characteristic, the dropouts as a group more often selected the negative response. The following descriptions are in terms of comparison -- dropouts versus graduates.

Table 2

Negative Response Pattern for the Dropouts

Immaturity
Rebellious attitude toward education
Escape
Anti-establishment attitude
Lack of achievement motivation
Anti-social behavior

The dropouts tended to be more immature than the graduate group. The younger the individual, the greater likelihood of attrition. Approximately 44 percent of the dropouts were 17 or under as compared to about 22 percent of the graduates. This reflects a 22 percent difference between the dropouts and graduates. At the age level 21 or older there was about a 13 percent difference indicating that the older the individual the greater the probability of his graduating from recruit training.

As a group, the dropouts tended to have a negative attitude toward education. They indicated that their parents or guardians were not too concerned about their educational attainment and were not pleased with their grades. The dropouts obtained less education, and received lower grades than the graduate group. More of the dropouts reported they cheated on tests in school, were suspended from school and/or were in disciplinary trouble, disagreed with their teachers, were more often truant from high school, and in general, disliked school.

The dropouts selected item alternatives which are interpreted as an effort on their part to escape from their current situation. More of the dropouts than the graduates were inclined not to like themselves in that if they could change they would like to be someone different from themselves. More of the dropouts stated that they liked getting drunk, and living in an orderly way bores them. They indicated that they have friends who have been busted for drugs and also they feel that the bad effects of marijuana and other drugs have been overemphasized.

The dropouts indicated a negative attitude toward authority and discipline and a general disregard for law and order. The dropouts stated more often than the graduates that they drove their parents' car without their permission. They don't mind people breaking the rules, they participated in protests and peace demonstrations, and would like to make a lot of changes in the laws of this country. They also feel it's OK to steal sometimes and used to steal when younger. They like things considered bad, forbidden, or even dangerous. They describe themselves as not usually very law abiding and feel oppressed by laws. In addition, they feel that the Navy dehumanizes people.

With respect to lack of achievement motivation, as a group they tended to daydream more in school and stated that their friends felt they were lazy. Substantially more were fired from jobs. They tended to put things off and have an "easy come, easy go" attitude. They also indicated they avoid hard work if possible.

Again, as a group they exhibited anti-social behavior and a negative outlook on life. The dropouts are somewhat uncooperative, easily aroused to anger and grouchy. They tend to act without thinking and have a disregard for others' problems, feelings, and/or property. They indicated that they feel they have had more than their share of bad luck in life and also indicated that life usually hands them a pretty raw deal.

The findings to date show promise of the DBI as a possible screen against some facets of delinquent behavior. However, the findings at this point in time are not yet conducive enough to warrant an outright recommendation for use. In addition, it would not be advisable to use the DBI scores by themselves for screening. The DBI is not envisioned as a "go-no-go" enlistment standard but rather as a management tool with which recruiters could, with other available information, screen out high risk applicants. The DBI's usefulness would probably be best utilized as an additional variable to the Odds for Effectiveness (OFE) scores or whatever primary screening instruments are in effect. The OFE table was developed to aid recruiters in estimating the odds that an individual applicant will effectively complete his initial period of active duty obligation and be recommended for reenlistment. The table gives chances in 100 that an applicant will render effective service as a function of his AFQT score (equivalent AFQT from SBTB or BTB), and years of education, expulsions and/or suspensions from school and the number of arrests, excluding traffic violations. At a minimum, identified high risk applicants could be more throughly interviewed by recruiters to determine their military suitability. The instrument could also be useful in detecting individuals who may have serious adjustment and behavioral problems early in their tour of duty.

REFERENCES

Yellen, T.M.I. Validation of the Delinquent Behavior Inventory as a predictor of basic training attrition. San Diego: U.S. Navy Personnel Research and Development Center, August 1975. (Technical Report 76-3)

Table 1. Distribution of DBI Scores in Relation to Graduation/Attritee Categories

	Graduates			Attritees		
Scores	Frequency	Cumulative Frequency	Cumulative Percent	Frequency	Cumulative Frequency	Cumulative Percent
29.55 to 32.99	4	2,384	100.	0	101	100.
26.10 to 29.54	41	2,380	.988	0	101	100.
22.65 to 26.09	212	2,339	.981	3	101	100.
19.20 to 22.64	277	2,127	.822	2	93	.970
15.75 to 19.19	435	1,850	.776	3	96	.95
12.30 to 15.74	265	1,415	.593	5	93	.921
8.85 to 12.29	341	1,150	.482	7	88	.871
5.40 to 8.84	182	809	.337	8	81	.802
1.95 to 5.39	204	627	.263	11	73	.723
1.94 to -1.50	123	423	.177	10	62	.614
-1.51 to -4.95	86	300	.126	9	52	.515
-4.96 to -8.40	100	214	.090	5	43	.426
-8.41 to -11.85	32	114	.048	11	38	.376
-11.86 to -15.30	45	82	.034	10	27	.267
-15.31 to -18.75	16	37	.016	5	17	.168
-18.76 to -22.20	15	21	.009	7	12	.119
-22.21 to -25.65	5	6	.000	1	5	.050
-25.66 to -29.10	1	1	.000	4	4	.040

A SURVEY OF THE DEVELOPMENTAL RESEARCH FOR PROJECT CONTRACT (COMPUTERIZED NAVY TECHNIQUES FOR RECRUIT ASSIGNMENT, COUNSELING, AND TESTING)*

by

Bernard A. Rafacz

Navy Personnel Research and Development Center
San Diego, California 92152

INTRODUCTION

As with the other branches of the Armed Services, the U. S. Navy is seeking ways of enlisting sufficient numbers of qualified individuals to satisfy projected manpower requirements. To meet this goal, under the greater demands placed on recruiting efforts, the Navy has adopted several policies as a means of inducing men and women to join its team. These efforts include an increase in the advertising effort, an increase in the number of recruiting stations and personnel, the reorganization of recruiting under a central agency, i.e., the Navy Recruiting Command (NRC), and the guaranteeing of prospective recruits (whom we shall refer to as applicants) their choice of training in an occupational specialty school program.

Even with these innovations, a Navy recruit applicant is required to decide upon one of a large number of unknown and unfamiliar ratings which he feels will yield maximum satisfaction and realization of his potential while serving in the Navy. On the other hand, the Navy has manpower requirements to meet in many different job areas, usually set within strict time constraints. Finally, in the middle of these requirements and frequently conflicting goals, is the Navy recruiter/classifier. He must be very much a "salesman" when counseling an applicant; that is, trying to give the applicant what he desires and at the same time meeting the quotas established by the recruiting command. Furthermore, he must "sell" the applicant a training program while being uncertain of the applicant's abilities and interests. In effect he is expected to counsel the applicant into a training program while having a very limited knowledge of the quality of the total applicant manpower pool available to the Navy for a particular point in time. The recruiter (and applicant) must decide upon a choice of training on the basis of information obtained at the recruiting station. Each recruiter must match this information for a recruit applicant with the skills and behaviors required by the various Navy jobs. Needless to say, in this counseling process exists the problem of how uniformly recruiters interpret and dispense information to Navy applicants prior to enlistment.

---

* The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

Present indications of recruiting policies suggest an increased use of assignment commitments prior to enlistment. Approximately 50% of the present recruit population are being guaranteed an occupational specialty school program. Because such assignments must be made within a relatively short time frame, dynamic and time flexible models must be developed. The static, time-locked algorithms presently utilized by the Navy to assign a pool of recruits cannot be utilized for the assignment process that will be required. Furthermore, it is becoming more and more critical to consider the recruit candidate's interests, background, etc. in addition to aptitudes, in an optimization function. It is envisioned that a nonlinear dynamic model rather than the present linear model will provide the solution to these foreseen problem areas involving testing, counseling, and assignment. As such, it is necessary to test the feasibility of combining the occupational choice aspect of vocational guidance systems with the one-by-one assignment system into a technology that can be adapted into the present, and projected, recruiting environment.

From a research and development point of view, beyond the practical considerations mentioned above, there exists the problem of an appropriate vehicle within the present-day recruiting environment for implementation of research developed counseling tools. An example of this is the Navy Vocational Interest Inventory (NVII) which presently applies to fifteen Navy ratings. However, because of the time delay occasioned by mailing and to score the instrument, its full utility has not been realized. The scoring problem has been resolved by the development of a procedure (Rafacz, 1975) that would enable recruiters to access a computer facility (via a remote terminal) for the purpose of administering the NVII.

The format of this paper is divided into a number of general topical areas. The first area of discussion is concerned with related research efforts from the military and civilian communities. The second topical area deals with an ADP equipment prototype that is expected to best serve as the monitoring system for the Navy's recruiting environment. This discussion is followed by an outline of the research efforts to be pursued under project CONTRACT. An implementation plan for the prototype system over successive stages is then outlined. Finally, various ramifications of the monitoring system are presented to include possible areas in which the system will be of benefit to the Navy.

COMPUTERIZED COUNSELING SYSTEMS

Military Systems

NAVY

In the military setting, several assignment and/or counseling systems have been developed and implemented or are being developed for possible implementation that have a direct bearing on the research efforts of this project. Present Navy systems include the Navy's Recruit Allocation Control System (RACS). RACS is essentially a telephonic inquiry system utilizing WATTS lines connecting recruiters with the Navy Recruiting Command. This system accounts for regular Navy accessions who have been guaranteed service school training spaces (Class A-school seats) upon completion of recruit training. Approximately 65,000 accessions, who have been guaranteed Class A-school seats, per year are processed through the system. RACS produces management reports by means of an off-line computer support system. Basically, RACS is a system that facilitates the Navy's program of guaranteeing service school training spaces and occupational specialty school programs to Navy accessions at the recruiting station level. Necessarily such guarantees are on a first-come, first-service basis. In this sense, RACS is analogous to an airline reservation system.

AIR FORCE

The Air Force's presently utilized Accessions Control Center (ACC) is almost identical to the Navy's RACS. Essentially, ACC consists of a WATTS line communication link between field recruiters and the Military Personnel Center's Procurement Management Information System (PROMIS). ACC controllers are stationed at three round tables to which are attached large rotary files of punched cards. Each card corresponds to a training school slot for a guaranteed training enlistment. The system has the capability of reserving training school slots and effecting cancellations, but not on a real-time basis.

The Air Force has been conducting research on an Advanced Personnel Procurement System (APPS) that addresses alternatives to ACC. Hawkins, Crow, and Haltman (1974) have prepared a report on six different configurations of an APPS for the Air Force. The most advanced configuration of computer software and hardware evaluated consists of teletype terminals operating at 110 baud located in each of the Air Force's 997 recruiting offices. The General Science Corporation (GSC) INFONET system would connect these terminals to a centralized computer system. This configuration will require Air Force recruits to enter an applicant's qualifications and desired entry dates for analyses into a computer system. By a person-job match algorithm the computer system will examine all job vacancies available and select those training slots or aptitude areas in which the individual is best qualified to train, and best match the prospect's abilities and interests. The most sophisticated configuration evaluated in the report is essentially a configuration that interacts with a recruiter. The other five configurations include the location of terminals in recruiting sectors, detachments, or groups, at AFEES, or upgrading the presently utilized ACC.

ARMY

The Army on the other hand has been working on its project REQUEST. This project involves an interactive real-time sharing system utilizing the INFONET communications network. Software for the system was developed by Systems Automation Corporation. The project has terminals located in the Guidance Counselor Offices in each of the 74 Armed Forces Examining and Entrance Stations (AFEES). In addition to the terminals at the AFEES, the Army is assigning some to mobile recruiting teams and increasing the number of terminals to over 100. The GSA contracted Computer Sciences Corporation (CSC) INFONET system to provide nationwide linkages from AFEES to a UNIVAC 1108 computer complex near Los Angeles.

REQUEST is presently being expanded to include all enlistments and allocation of all Army training spaces. However, it is still basically an airline reservation system and all reservations for school training spaces are made on a first-come, first-served basis. The counselor at AFEES is the only interface between the Army applicant and the REQUEST system. The counselor enters all information on a prospect's physical profile, test scores, desired entry dates and desired occupational area. REQUEST matches this information against requirements for various occupational areas and the available quotas. Once the search is completed the best matches are displayed and the counselor and applicant decide on an occupational area.

A related research effort utilizing an individual/computer effort is the Officer Career Counseling research being conducted by the U. S. Army Research Institute (ARI) for the Behavioral and Social Sciences. The researchers at ARI involved in this project have developed an Experimental Computer Aided Career Counseling System (ECACCS) which when fully developed would utilize portable terminals accessing a central computer. After several years work they have developed an experimental system that allows Army officers to develop long-range career planning strategies and receive instruction (on a real-time basis) on current career paths and policies. Of particular interest are the gaming models (simulation of officer careers) and the interactive counseling dialogues developed for their purposes.

Civilian Developments

There are also systems for assignment and/or counseling within the civilian environment that are related to research efforts of this project.

A significant contribution is a computer-assisted occupational selection system (Impellitteri, 1968), utilizing an IBM 1050 computer with remote terminal facilities, tape recorder-player, and slide projector. This system stores student aptitude score profiles and allows for student interaction with the system, providing job information requested as well as amplifying information on a particular occupation via remote facilities (slide projector, etc.). Field test results indicate that the system is effective and has utility in assigning youth in the exploration of their occupational opportunities.

Cogswell and Estavan (1965) of Systems Development Corporation, in conjunction with the University of Oregon and others, have developed a computer-assisted counseling system utilizing a Philco 2000 computer and based on an inferred model of a counselor's decision processes. Validity studies in which computer output was compared with output of a model counselor showed about 75% agreement between the system and counselor.

Perhaps the most impressive computer based counseling system found in the civilian sphere is that developed by the Educational Testing Service (ETS) researchers (Katz, Chapman, Godwin, 1972) called System of Interactive Guidance and Information (SIGI). Oriented primarily toward junior college students, a student interacts with the system in such a manner that makes it possible to examine his own values, obtain and use relevant information, interpret predictive data and to formulate plans. SIGI very closely satisfies the proposed automated occupational counseling and applicant inquiry system to be developed by NPRDC for use within the Navy recruiting environment.

AN ADP EQUIPMENT PROTOTYPE FOR THE RECRUITING ENVIRONMENT

It is clear that the impetus of the previous sections demonstrate that support of both a research and applied nature can be accomplished by a system which at the same time is capable of performing all testing on an applicant (i.e., is a hands-off data gathering system), has rapid turnaround capability (i.e., it is essentially a real-time system), is uncannily accurate, is able to interpret vast amounts of information in seconds, has rapid entry/retrieval of large data storage areas, is relatively easy to communicate with, and is capable of performing complicated mathematical routines. It is the purpose of this section to present a configuration of automatic data processing (ADP) equipment that may be utilized within the present-day recruiting environment to provide such support.

FINDS Components

The ADP system outlined herein is an exact prototype of a system which could be eventually placed into an operating environment following the development of needed research and required software packages. The proposed collection of computer hardware and software will be identified by the acronym FINDS (Functionally Interactive Navy Decision System). Figure 1 is provided as a guide for describing the various components of FINDS and the general functions that each unit will perform. Figure 2 is a more detailed explanation of the component parts of the system as found in one recruiting district.[1]

Central Processor

The main unit of FINDS is the central processor. The central processor is really a minicomputer, i.e., it is capable of being programmed utilizing a language similar to FORTRAN, BASIC, COBOL, etc. The memory of the central processor will store a programmer's instructional logic for execution by a recruiter/applicant at some point in time. The central processor's primary function is to provide the logic (i.e., software packages) for such functions as: (1) retrieving test, interest, and demographic items from a storage media and presenting those items to applicants on terminals; (2) checking of the responses as to whether they are of the correct type (not duplicate responses to same item and not an unacceptable alternative); (3) storing test responses and applicant background data on data storage devices; (4) performing small scale mathematical routines (scoring tests or developing occupational area scores on interest measures); (5) satisfying user's (applicant/recruiter) inquiries for specific information about the Navy; and (6) maintaining interactive career dialogues.

CRT Terminals

The second element of the system is the terminals. The terminals are connected to the central processor by a communications line (i.e., telephone line). By virtue of this connection they are able to utilize the logic capabilities of the central processor. Therefore, by communicating with the central processor each applicant can assess tests or cure the central processor for specific information.

Host Computer System

A critical element of FINDS is a host computer system. Depending upon geographical location and availability of rental time, one computer

---

[1] The term "district" is employed here to represent a collection or group of recruiting stations which, because of their relative location, represent minimum communication costs when communicating with a minicomputer.

Figure 1
Main Components of FINDS

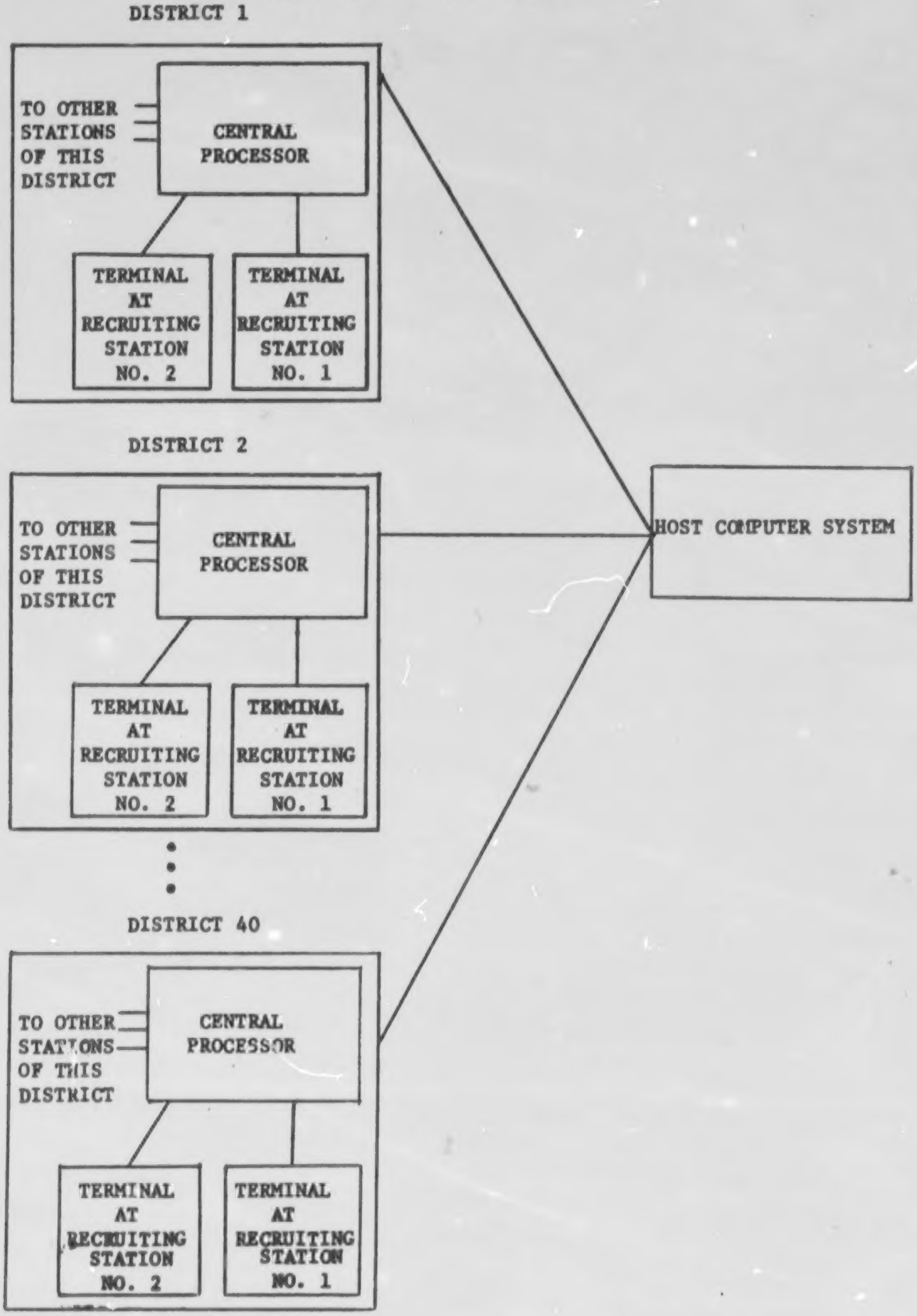
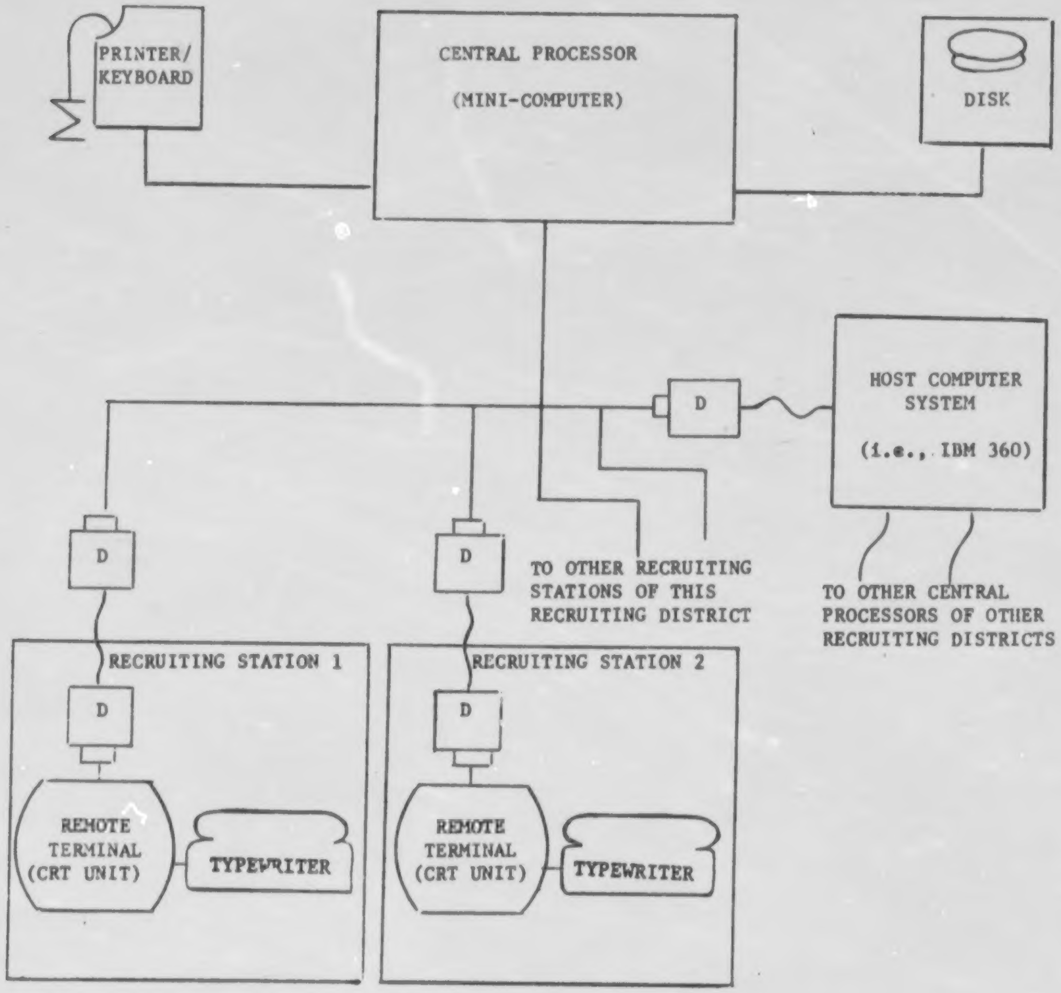

DISTRICT 1
TO OTHER STATIONS OF THIS DISTRICT
CENTRAL PROCESSOR
TERMINAL AT RECRUITING STATION NO. 2
TERMINAL AT RECRUITING STATION NO. 1
DISTRICT 2
TO OTHER STATIONS OF THIS DISTRICT
CENTRAL PROCESSOR
TERMINAL AT RECRUITING STATION NO. 2
TERMINAL AT RECRUITING STATION NO. 1
DISTRICT 40
TO OTHER STATIONS OF THIS DISTRICT
CENTRAL PROCESSOR
TERMINAL AT RECRUITING STATION NO. 2
TERMINAL AT RECRUITING STATION NO. 1
HOST COMPUTER SYSTEM

Figure 2

A Central Processor Control Unit as Setup in Some Recruiting District


PRINTER/ KEYBOARD
CENTRAL PROCESSOR
(MINI-COMPUTER)
DISK
HOST COMPUTER SYSTEM
(i.e., IBM 360)
D
D
D
TO OTHER RECRUITING STATIONS OF THIS RECRUITING DISTRICT
TO OTHER CENTRAL PROCESSORS OF OTHER RECRUITING DISTRICTS
RECRUITING STATION 1
D
REMOTE TERMINAL (CRT UNIT)
TYPEWRITER
RECRUITING STATION 2
D
REMOTE TERMINAL (CRT UNIT)
TYPEWRITER

D - ILLUSTRATES A DATASET BETWEEN THE COMPONENTS
- ILLUSTRATES A TELEPHONE LINE CONNECTION BETWEEN THE COMPONENTS

site could service more than one district. A host computer would be required for executing the computer programs which are beyond the capabilities of the central processor. The central processor can communicate with the host computer for the purpose of transmitting applicant data for subsequent large scale data analysis.

Selection of FINDS

NPRDC researchers have evaluated various alternative computer hardware and software configurations that could eventually be utilized as a monitoring system of the Navy's recruiting environment. A prototype system has been identified and procedures for procuring the candidate system have been initiated. The proposed prototype system essentially consists in a minicomputer, disk storage area, and several terminals.

The selected prototype system was identified mainly with respect to the way in which the interaction among the various components of the system is to be accomplished. In this regard, it was foreseen as a necessary condition for selection that the central processor be capable of controlling a time-sharing environment while maintaining, without performance degradation, communication with a host computer system. Futhermore, it was required that the selected system include the software for a sophisticated data base management system and support FORTRAN programming. This will enable project researchers to avoid the necessity of developing the software for several critical research requirement areas within project CONTRACT.

Project researchers are continuing in their evaluation of various alternative computer systems. In this regard, a cost-effectiveness and time-savings trade-off of the selected system has been initiated.

RESEARCH AREAS

The following sections elaborate on the research topics to be explored under project CONTRACT. These areas are seen as necessary to obtain a Navy recruiting monitoring system that can unify the areas of testing, counseling, and assignment into a systematic and timely procedure within the present and anticipated recruiting environment.

Evaluating Procedures for the Acquisition, Retrieval, and Storage of Testing Information

At present, Navy applicants are administered several tests and forms which elicit information concerning their aptitudes, interests, biographical, and demographic information. Scores from these instruments are used primarily to determine an applicant's eligibility for training in occupational specialty programs. Indications are that present aptitude testing instruments will be replaced in the near future by a unified battery of tests, Armed Services Vocational Aptitude Battery (ASVAB). Utilizing ASVAB information, research is presently being conducted at NPRDC concerning an applicant's success, performance, or satisfaction in any of the many possible Navy jobs and training schools.

The research in this area relative to project CONTRACT will be involved with developing appropriate computer dialogues between an applicant and FINDS to obtain a data file or profile on an applicant. The profile should include such information as aptitude test measures and information concerning the applicant's interests and attitudes, career objectives, performance capability, and, of course, biographic and demographic information. Once this set of information has been obtained on an applicant, it will be used to derive a mathematical (regression) estimate of an applicant's future performance and success in candidate Navy training schools or programs and job areas. Necessarily the research must include the identification of those techniques (adaptable to a computerized monitoring system) which could minimize the testing difficulties of length of a test/questionnaire and total examination time.

Essentially, then the testing phase of the research effort involves the identification and evaluation of sample configurations of administering various tests and collecting data via CRT terminals. This will include the development of: (1) appropriate test dialogues between the applicant and FINDS; (2) effective scoring procedures; and (3) providing meaningful estimates of applicant success in various Navy ratings or occupational specialty school programs. This will relieve recruiters of the problems they presently have in scoring some test instruments (e.g., the Navy Vocational Interest Inventory (NVII)) and also provide a uniformity in the interpretation of applicant success as a function of test, interest, and biographic variables. Inherent in this phase is the development of the schema and computer software package for extracting information from applicants and recording it in a centralized location so that an information retrieval system can extract the appropriate data when needed. In essence, the research will be concerned with the establishment of a Navy recruit applicant data base and management system.

Occupational Counseling and Applicant Inquiry

Computerized assisted counseling in application will probably consist of a three-fold interactive dialogue among the applicant, recruiter, and computer system (via CRT's). The result is a dialogue of the characteristics and opportunities available to the applicant if he selected any one of the occupational specialty programs for which he qualifies. Minicomputers controlling the computer conversation on the CRT's could access a number of data banks and thus be capable of answering most of the questions applicants normally ask recruiters about the

Navy or their possible work areas. Interactive counseling dialogues could then be utilized to simulate possible career paths for a Navy applicant.

The research effort in the counseling phase of project CONTRACT will be concerned with developing the appropriate computer dialogues for recruit applicant long-range career planning. This means the development of real-time interactive dialogues between the applicant (his preferences and qualifications) and FINDS (its updated account of current career possibilities the Navy policies). Furthermore, it will be necessary to develop the software relative to the dynamic decision process that will occur. The applicant will be changing his preferences while FINDS is providing him information.

As such the research effort will address the development of means of utilizing the information derived from applicant testing to counsel the individual into that training program (and associated job area or rating) which best matches the applicant's career aspirations and qualifications and the Navy's then current needs and policies. In this sense the research effort will involve developing a real-time interactive computerized dialogue that controls a decision-making process and is capable of identifying the career progression patterns that would be of greatest benefit to the applicant and the Navy.

One aspect of the research within the area of recruit applicant counseling is the identification and modification of already existing information retrieval systems into a unified naval system that can evaluate applicant inquiries of Navy manpower statistics on a real-time basis. Of primary concern will be the adoption of those techniques into the present recruiting environment. The system will be unified in the sense that information derived from applicant testing results will be integrated into the counseling session. Another aspect of the research effort within this area is the modification of already-existing and implemented computerized counseling systems for possible application within the recruiting environment. Appropriate testing designs for evaluating various systems will be developed, borrowing from similar research efforts within the civilian community (e.g., Bohm, 1970). In an analogous manner, the advice and consultation of individuals who have developed or worked on such systems is being solicited. It is expected that these efforts will enable NPRDC to avoid "recreating the wheel" and focus more upon the possible application of already developed systems into the Navy.

The Navy will benefit considerably from research within the computerized counseling area in that the proposed monitoring system will provide standarized occupational information to the recruit applicants and present the current career possibilities available to the applicant. Thereby, information bias introduced to applicants by recruiters can be minimized, or, at least, better controlled. This will free the recruiter/classifier from the necessity of having a timely and encompassing knowledge of the Navy and at the same time be of maximal benefit to the applicant in selecting a job or training school of his choice.

Optimal Applicant Assignment

In the present recruiting climate an applicant is offered a training program if he qualifies and desires to be enlisted into that program; a first-come, first-served assignment policy prevails. However, this policy necessarily restricts more qualified applicants from desired school assignments which have already been filled by applicants who arrived at recruiting stations earlier in a recruiting time frame. Necessarily, such a policy could never be optimal; i.e., maximize overall payoff to the Navy.

To clarify the issue, in the assignment problem considered here, it is not necessarily true that an applicant will be accepted by the Navy into a particular occupational specialty just because he qualifies and desires it. Rather he is accepted in that area because out of the pool of presently accepted recruits in that occupational specialty, together with the pool of projected recruit arrivals into the Navy recruiting system, he is the best possible recruit the Navy could have for the time the applicant arrives at a recruiting station. He is best in the sense that no other recruit (having arrived yet or not) can be expected to have a greater degree of success and be of more benefit to the Navy. This assignment problem is certainly in sharp contrast to the assignment technique utilized today in making individual assignments at the recruiting stations which are, as previously mentioned, made on a first-come, first-served basis. It is a reasonable hypothesis that the optimal assignment procedure should increase performance levels and satisfaction in each Navy job area.

Research efforts within the optimal applicant assignment area will be twofold. First, it will be necessary to develop a simulation of the present recruiting environment. This implies the development of a computer simulation model that includes as many variables as will be possible to be included in a Navy recruiting monitoring system if it were implemented at some future time. Inherent in this effort will be the identification of all parameters to be included in the model, the assumptions which have to be made, and what information is presently available on the recruit and applicant populations. Finally, this part of the effort will include a complete documentation of the simulation effort (e.g., flowcharts, definitions and descriptions of printout information). Secondly, (and concurrently with the development of the simulation effort) the research will involve the identification of candidate optimal assignment strategies and the means by which they are to be evaluated with the contexts of the simulation program. Included in this effort will be the identification of appropriate "payoff" or criterion matrices of future "applicant" success in the Navy. In this regard it will be necessary to develop a means of comparing various assignment policies (e.g., preference for assignment, expected A-school success) and, of course, the utilization of the appropriate statistical procedure for comparing the manpower policies that are being evaluated.

PLANS FOR IMPLEMENTATION OF CONTRACT

It is clear that the implementation effort underlying project CONTRACT is probably the most crucial aspect of the entire research effort. The project is concerned with the development of computerized procedures and a monitoring system which could serve as a vehicle for researcher developed instrumentation and counseling procedures. Implementation will essentially involve resolving any remaining practical utilization aspects of the proposed system.

As can well be imagined, implementation of the prototype system on a nationwide scale must occur over several phases. At present three distinct phases are viewed as necessary so that implementation on a nationwide scale can occur. Briefly these phases consist of:

<ins>Phase</ins> <ins>I</ins>: Development of a true prototype Navy recruit applicant counseling system. The research efforts will be concerned with developing procedures for implementing research developed counseling tools via the prototype system. It is expected that once the system is adequately refined, researchers will have a direct line to a field application area via the prototype system. Trial runs of the system will be conducted at various levels within the San Diego recruiting district.

<ins>Phase</ins> <ins>II</ins>: Development of a computerized counseling system for the Pacific Southwest area (includes the states of California, Arizona, Utah, and Nevada). This phase of the research effort will be primarily concerned with developing a network of computers to serve as the monitoring system for the approximate 140 recruiting stations in that area. The prototype system will be further refined to include the testing and evaluation of more sophisticated computerized counseling dialogue. It will also be possible to collect recruit applicant data and create a data base for the research efforts involved in further refining the prototype computer system (e.g., optimal assignment). The operational aspects of the monitoring system will be identified and further evaluated together with a clearer picture of the true trade-offs of the system in terms of cost benefits and over-all effectiveness.

<ins>Phase</ins> <ins>III</ins>: Nationwide Implementation Effort. It is not entirely clear at this time exactly what efforts will be required in this phase as much depends on the results of Phases I and II. Necessarily a cost

analysis of the entire system together with specific plans for incorporating the monitoring system into then existing Navy computer system will have to have been completed. In any event, refinement of the prototype will continue and revisions implemented as appropriate and feasible. It is anticipated that research efforts in this phase will involve developing procedures to enable Navy manpower planners to achieve a high level of accessibility of recruiting type data. Inherent to this effort will be the incorporation of the optimal assignment strategy (identified in the other phases), that is of maximal benefit to the Navy.

SUMMARY AND CONCLUSIONS

Project CONTRACT is concerned with developing computerized procedures and techniques for utilization by the Navy Recruiting Command at the recruiting station level within the areas of testing, counseling, and assignment. These techniques will ameliorate some of the problems recruiters and Navy manpower managers presently have in those areas. Obviously this effort will require the expertise of individuals involved in many disciplines, to include Operations Research, Statistics, Psychology, and Systems Analysis. Although the research effort will be initially concerned with utilizing already developed techniques, it will undoubtedly be necessary to develop new and innovative techniques in order to address the particular environment which concerns the project. Certainly many technological advances in systems analysis and psychological evaluation, to name a few, have been developed in related research efforts. Project CONTRACT will be concerned with these developments and will evaluate their merits within the context of the problem areas it addresses.

Of fundamental importance to project CONTRACT is the development of a prototype computer network that may be utilized by the Navy Recruiting Command to serve as the vehicle for monitoring the recruiting environment. This will result in a direct line for the implementation of research developed techniques for utilization with that environment. The project consumer will now have a means for assessing the merits of implementing a particular technique. Furthermore, he will also have an almost real-time procedure for assessing the status of the recruiting environment over many different levels and with respect to many possible factors.

Current efforts on the project are concentrated on the evaluation of several possible configurations of computer network systems that could serve the monitoring function. Once this issue has been resolved, a cost-effectiveness and manpower savings study will be initiated in order to better assess the merits of the proposed system.

In conclusion, the end products of project CONTRACT will enable the Navy Recruiting Command:

1. Guarantee optimal service school training spaces and occupational specialty school programs to Navy applicants at the recruiting station level, without recruiter intervention.

2. Perform testing of applicants over a number of instruments with scoring and analysis automatically provided.

3. Dispense information concerning Navy programs and policies in a more uniform fashion; minimize recruiter bias.

4. Provide any type of permanent record on an applicant (or sets of applicants) at any point within the Recruiting Command. This includes the automatic generation of summary statistics on the recruit applicant pool.

5. Automatically update recruiting manuals, guidelines, and instructions.

6. Simulate a Navy recruit applicant's career path as a function of his own unique qualifications.

These are but a few areas in which the research efforts of project CONTRACT may be employed. In fact, it is also a major consideration of the research effort to identify those areas in which possible end products of the overall research effort may be applied within the Navy.

REFERENCES

1. Bohm, Martin J., "Field Trial and Evaluation of a System," an article appearing in Computer Assisted Counseling, by Donald E. Super, Teachers College Press, Columbia Univ., pp 83-99, 1970.

2. Cogswell, J. F. and Estavan, D. P., "Explorations in Computer-Assisted Counseling," System Development Corporation, Santa Monica, CA, TM-2582/000/00, 1965.

3. Hawkins, M.D., Crow, R., and Haltman, H., Cost Estimates and Implications of Alternative Configurations of an Advanced Personnel Procurement System (APPS). Lackland Air Force Base: Air Force Human Resources Laboratory, AFHRL TR 74-82, 1974.

4. Impelliterri, J. T., "A Computer-Assisted Career Exploration System," National Business Educational Quarterly, 37, 1968, pp 47-51.

5. Katz, M., Chapman, W., and Godwin, W., "SIGI-A Computer-Based Aid to Career Decision-Making." Reprinted from EDUCOM Bulletin, Summer Issue, 1972, 6p.

6. Rafacz, B. A., *Manual for a Remote Terminal Computer Analysis of the Navy Vocational Interest Inventory*. San Diego: Navy Personnel Research and Development Center, NPRDC TR 75-1, 1975.

NEW CODAP PROGRAMS FOR ANALYZING TASK-FACTOR INFORMATION

By

Raymond E. Christal
Occupational and Manpower Research Division
Johnny J. Weissmuller
Computational Sciences Division
Air Force Human Resources Laboratory (AFSC)
Lackland AFB, Texas

I am making the assumption that most of your here are familiar with the CODAP analysis system, which is a comprehensive set of computer programs for analyzing and reporting occupational information collected with job inventories. The CODAP system presently contains approximately thirty-five main programs which are oriented toward producing task-level descriptions of the work performed by individuals and groups of individuals. The system also contains a number of programs for identifying and describing the types of jobs existing in an occupational category, and for describing the characteristics of individuals falling into special or job-type groups. Today, I would like to describe a new series of programs recently added to the CODAP system which are oriented toward analyzing and manipulating information describing work tasks, rather than jobs or persons. These programs fill an important gap by equipping the analyst with tools for addressing new problem areas.

Before describing the new programs, I should recognize that CODAP already contains some programs which could be considered as being task-oriented. For example, ASFACT is designed to summarize task-factor rating information provided by job incumbents, while GRPSUM can be applied to display the percent of members in various groups performing any given task. AVALUE and AVGPCT can be used to compute such indexes as the average grade level or time-in-service of individuals performing each work task in an inventory. Two other task-oriented programs in the UNIVAC 1108 version of CODAP are REXALL and TSKNDX. These will be described briefly, since they interact with the new programs now being released.

REXALL is a very powerful and flexible program for analyzing the inter-rater agreement among judges concerning task-factor ratings. This program also generates a task-factor weight deck for input into other CODAP programs. An example output from a REXALL run is given in Table 1. This table displays certain information concerning task-factor ratings provided by each of 38 raters participating in a recent task

Table 1. REXALL Rater Correlations with Sample Means--Raw Responses

Rater ID	r_{xy}	N Ratings	Mean	S.D.	T-Value
1	.735	413	3.57	1.06	19.12
2	.655	414	4.37	0.92	15.30
3	.777	414	3.77	0.96	21.79
4	.658	414	3.56	1.08	15.43
5	.636	414	4.16	1.58	14.55
6	.697	414	2.96	1.00	17.16
7	-.474	413	4.59	1.46	-9.48
8	.791	414	4.54	0.88	22.84
9	.730	414	4.34	1.06	18.87
10	.690	414	5.09	.72	16.84
11	.031	414	4.04	1.01	0.55
12	.789	414	3.93	1.14	22.74
13	.614	414	4.07	1.16	13.73
14	.819	414	4.66	1.51	25.21
15	.639	414	3.82	1.29	14.68
•	•	•	•	•	•
•	•	•	•	•	•
•	•	•	•	•	•
38	.805	414	4.49	0.96	23.95

difficulty analysis. The program first computes the mean rating for each task across all raters; then it correlates the ratings provided by each particular rater with the vector of mean ratings obtained from the total sample. These correlations range from .61 to .81. The values provided by rater #7 correlate -.47 with the ratings provided by the complete sample. He used the rating scale upside down. This is not unusual. We have analyzed task-factor information from hundreds of raters on a variety of task factors, and we find that about 5% of the participants reverse the rating scale. Rater #11 is "non-cooperative." It is inconceivable that there would be *no* agreement between a supervisor and his peers concerning the relative difficulty of tasks in his occupation. This is an example of an individual who simply "filled in the squares" with random numbers. Both of these individuals contribute "error" variance to the mean factor vector, and, at AFHRL, we would elect to discard them from the data set. Another source of error is reflected in column 4, which reports the mean of ratings provided by each rater. In this particular study, the raters were asked to rate the *relative* difficulties of tasks in their occupation using a 7-point scale. The values in column 4 reveal that some raters tended to use the upper part of the scale, while others tended to use the lower part. For example, rater #6 had a mean of 2.96, while rater #10 had a mean of 5.09. Since all ratings are *relative*, there is no reason why the mean rating across all tasks should vary significantly from rater to rater. Such differences in means introduce within-task variance, which is treated as error variance in computation of the inter-rater agreement coefficient. REXALL provides a method of eliminating such error variance by standardizing ratings.

Table 2 presents summary data associated with the first REXALL run in the example. Notice that the reliability coefficient for the means based upon all 38 raters is .939, which is not bad, considering that: (a) two raters were included who reversed the rating scale; (b) one rater was non-cooperative; and (c) the ratings were not standardized. The sample value is only .288, which can be loosely defined as indicating the reliability for a single rater. It is this value which is entered into the Spearman-Brown prophecy formula to estimate the reliability of mean ratings based upon a sample of any given size.

You will notice that REXALL provides a printout of all data entering into the computation of the reliability coefficient. It also indicates, among other things: the average number of ratings per task; number of raters; number of tasks; total number of task ratings; percent of possible ratings; mean and standard deviation of all ratings; mean and standard

Table 2. Inter-Rater Reliability Based on Raw Ratings

RKK	=	.93892	Reliability for These K Raters
R11	=	.28806	Reliability for a Single Rater
F11	=	16.3724	F-Ratio (BTMS/WMS)
DFB	=	313.	Degrees of Freedom (Between, N1)
DFW	=	11616.	Degrees of Freedom (Within, N2)
DFT	=	11929.	Degrees of Freedom (Total)
C	=	199970.	Correction Factor for Sum of Sqs
BSS	=	6560.	Between Sum of Squares
WSS	=	14871.	Within Sum of Squares
TSS	=	21431.	Total Sum of Squares
BTMS	=	20.96	Between Tasks Mean Squares
WMS	=	1.28	Within Tasks Mean Squares
K	=	37.99	Average Numbers of Raters per Task
NRATER	=	38.	Total Number of Raters
NTASK	=	314.	Number of Tasks to be Rated
NZTASK	-	314.	Number of Tasks Actually Rated
NRESP	=	11930.	Total Number of Ratings
PCTRSP	=	99.98	Percent of Total Possible Ratings
TOTMN	=	4.0941.	Mean of All Task Ratings
TOTSD	=	1.3403	Standard Deviation of All Ratings
AVEMN	=	4.0942	Mean of All Task Means
AVESD	=	1.1061	Mean of Task Standard Deviations
SDMN	=	7416.	Standard Deviation of Task Means
SDSD	=	.1518	Standard Deviation of Standard Deviations

deviation of task means; and mean and standard deviation of task standard deviations.

Normally, REXALL is executed at least twice to process data for a particular factor in a study. First, it is run on raw data in order to identify non-cooperative and upside down raters for discard. In the second run, the values from the retained raters are standardized, unless an absolute scale is being used. Table 3 presents the summary information for the second run of the example data. Notice that three raters were discarded, leaving 35 raters in the run. The raw scores for each rater and for the means across all raters were standardized to a mean of 5.00 and a standard deviation of 1.00. These actions raised the reliability of the mean ratings from .939 to .973, even though the number of raters was reduced. At the same time, the sample value was raised from .288 to .503.

One of the outputs from the REXALL is a weight deck containing the mean rated value for each task in the job inventory on the factor in question. This weight deck is in proper format for direct input into other CODAP programs. For example, mean task difficulty weights might be entered into the VARGEN program to compute the average difficulty of tasks performed by each worker in a sample. Normally, however, the weight deck is first entered into the TSKNDX program in order to display the mean ratings of each task, along with other relevant information. Table 4 presents an abbreviated output from a TSKNDX run for tasks in the Electrical Power Production career ladder. In this specialty, draining condensate from air receiver tanks is considered to be the simplest task, while repairing electronic governors is considered to be the most complex. The TSKNDX program is flexible. At the requestor's option, it will publish data in task order or from low to high or high to low on mean factor ratings. It also presents other data of interest, such as the number of workers in a sample of interest performing each task, as well as the percent of time being spent on each task.

ASFACT, AVALUE, AVGPCT, REXALL, and TSKNDX are task-oriented programs which have been available for a number of years in the IBM 7040 and UNIVAC 1108 versions of CODAP. All five of these programs are currently coded in FORTRAN V, and could easily be converted for execution on other computers. While these programs are all very useful and efficient, recent studies in job evaluation and curricula development have required analyses and report formats of task-factor data which could not be serviced by existing CODAP programs. For this reason, a new series of task-oriented programs has been written to fill the gap. These are enumerated in Table 5.

Table 3. Inter-Rater Reliability Based on Adjusted Ratings

RKK = .97251 Reliability for These K Raters

R11 = .50274 Reliability for a Single Rater

F11 = 36.3827 F-Ratio (BTMS/WMS)

DFB = 313. Degrees of Freedom (Between, N1)

DFW = 10641. Degrees of Freedom (Within, N2)

DFT = 10954. Degrees of Freedom (Total)

C = 274728. Correction Factor for Sum of Sqs

BSS = 5681. Between Sum of Squares

WSS = 5308. Within Sum of Squares

TSS = 10989. Total Sum of Squares

BTMS = 18.15 Between Tasks Mean Squares

WMS = .50 Within Tasks Mean Squares

K = 35.00 Average Number of Raters per Task

NRATER = 35. Total Number of Raters

NTASK = 314. Number of Tasks to be Rated

NZTASK = 314. Number of Tasks Actually Rated

NRESP = 10989. Total Number of Ratings

PCTRSP = 99.99 Percent of Total Possible Ratings

TOTMN = 5.0000 Mean of All Task Ratings

TOTSD = 1.0000 Standard Deviation of All Ratings

AVEMN = 5.0000 Mean of All Task Means

AVESD = .6847 Mean of Task Standard Deviations

SDMN = 1.0000 Standard Deviation of Task Means

SDSD = 1.0518 Standard Deviation of Standard Deviations

Table 4. CODAP: Task Difficulty Index (TSKDIF) for Electrical Power Production (N=530)

D	TSK	Duty/Task Titles	Task Sequence Number	Task Difficulty Index	Number of Members Performing	Percent of Members Performing	Average Percent Time Spent by Members Performing
I	10	Drain Condensate from Air Receiver Tanks	1	2.15	269	50.75	1.49
.	.	. . .	.	.	.	.	.
.	.	. . .	.	.	.	.	.
H	3	Adjust Engine Controls After Starting	119	3.82	352	66.42	1.77
L	36	Take Firing Pressure Readings	120	3.82	170	32.08	1.03
M	21	Replace Water Circulating Pumps and Valves	121	3.82	69	13.02	0.67
.	.	. . .	.	.	.	.	.
.	.	. . .	.	.	.	.	.
H	5	Compute Power Produced by Generator	198	4.22	208	39.25	1.37
M	2	Adjust High Water Temp. Alarms & Shut Down Equip.	199	4.22	147	27.74	1.08
N	6	Correct Causes of Brush Sparking or Arcing	200	4.22	155	29.25	1.02
.	.	. . .	.	.	.	.	.
.	.	. . .	.	.	.	.	.
J	8	Determine Cause of High or Low Oil Pressure	300	4.60	214	40.38	0.97
L	29	Replace Piston and Connecting Rod Assembly	301	4.60	108	20.38	0.76
L	32	Replace Rod and Main Bearings	302	4.60	95	17.92	0.72
.	.	. . .	.	.	.	.	.
.	.	. . .	.	.	.	.	.
P	4	Analyze Static Battery Charging Diagrams	407	5.38	41	7.74	1.06
M	4	Determine Cause of Hydraulic Governor Malfunctions	408	5.42	110	20.75	0.85
O	1	Adjust Automatic Start Systems	409	5.42	103	19.43	0.92
.	.	. . .	.	.	.	.	.
.	.	. . .	.	.	.	.	.
O	15	Inspect, Replace, or Repair Governor Computer	434	5.85	31	5.85	0.51
O	20	Repair or Replace Transistorized Switching Devices	435	5.85	38	7.17	0.63
P	6	Analyze Transistorized Switching Devices	436	5.88	22	4.15	0.80
.	.	. . .	.	.	.	.	.
.	.	. . .	.	.	.	.	.
M	14	Repair Electronic Governors	444	6.27	22	4.15	0.48

Table 5. New CODAP Programs for Analyzing and Displaying Task-Oriented Data

TSKFAC - Adds factor weight vectors to JOB DESCRIPTION file.

FACSTD - Generates file of rater background information and raw ratings.

FACSPC - Computes, displays, and saves mean factor ratings for identified groups of raters.

COMGEN - Generates and saves factors and factor composites through user-specified operations.

TSKCOR - Extracts vectors from JOB DESCRIPTION file, computes inter-correlations and user-specified regression equations.

PREFAC - Applies equation weights developed by TSKCOR, or CORREG, and produces predictor score vector.

TSKCAT - Generates user-specified vector of 1's and 0's to identify tasks of interest. Provides input to TSKFAC.

FACSUM - Flexible program for selecting, sorting, and printing task-factor information. Computes new vectors based on differences between vectors, maximums or minimums of vector sets, cumulative percentages, and categories of tasks developed by six arithmetic operators.

The first program is called TSKFAC. It is used to create new records on the JOB DESCRIPTION file which represent values for task factors. Decks containing factor values may be keypunched. However, such decks are normally generated by other CODAP programs, such as AVALUE, AVGPCT, REXALL, TSKGRP, or TSKNDX. The TSKFAC program produces Request cards, which can serve as control cards in AUTOJT, MTXPRT, FACSUM, and PRIJOB. Several factors can be written on the JOB DESCRIPTION file with a single execution of TSKFAC.

The second program, labeled FACSTD, serves the same function as the INPSTD does for job survey raw data in the present CODAP system, except it is used to create a Rater History file instead of a worker history file. Input to this program is the SETCHKed raw data file, which contains rating and history information. Input also includes Duty/Task title cards and History variable definitions. In the UNIVAC 1108 version, FACSTD will accept 20,000 raters (a highly unlikely encounter), 1700 task ratings, 1000 history variables, and 26 duty variables.

FACSPC calculates and publishes raw and standardized factor means for special groups of raters whose membership is defined in terms of some combination of background or computed variables. The mean ratings may be reported in high to low sequence on mean values, or on the percent of members performing, or in duty-task sequence. An option is provided for writing the output on the REPORT and/or JOB DESCRIPTION files. In some ways, FACSPC parallels the JOBSPC program, except that it applies to factor ratings instead of job descriptions.

COMGEN is a new program which primarily is used to generate composite factors in accordance with user specifications. However, this program is very general-purpose, operating on factor data in somewhat the same manner as PROGEN currently operates on history and task data. The COMGEN program generates coding from user-supplied verbs and FORTRAN statements to perform various mathematical operations on vectors from the JOB DESCRIPTION file.

TSKCOR extracts up to 100 vectors from a CODAP JOB DESCRIPTION file and computes matrices of intercorrelations as well as user-specified regression problems. The correlation matrices may be saved on tape for future use by CORREG.

PREFAC accepts equation weights developed by TSKCOR or CORREG and produces a predicted score deck which is ready for input by TSKFAC.

TSKCAT is a simple program which translates card decks containing task numbers to card decks with the identified tasks represented as "1's" in a NTASK--long string (vector) of 1's and 0's. The output deck is designed for input into TSKFAC and other programs for later use as an operator vector. For example, it can be used to suppress information in a FACSUM report.

FACSUM is the most flexible program in the new task-oriented package. This program allows the user to print any of the vectors on the JOB DESCRIPTION file. However, during the same run, the user can generate and publish new vectors representing differences between two vectors; the maximums or minimums of sets of vectors; cumulative percentages; or categories of tasks using any one of six arithmetic operators. Inclusion of sequence numbering, blank columns, sorting, and print suppression are under user control.

Now that the new programs have been enumerated, let's go back through and add a few details. First, it should be recognized that all of the new programs are now an integral part of the CODAP system. They interact with each other and with CODAP programs in the previous version. FACSUM can extract, manipulate, and display any vectors on the JOB DESCRIPTION file, including the "percent performing," "time spent by those performing," and "time spent by total group" vectors associated with any previously computed job description. Task-factor vectors generated by five previously available CODAP programs can be written on the JOB DESCRIPTION file by the TSKFAC program. Task factors placed on the JOB DESCRIPTION file can be manipulated by five previously available CODAP programs. Output from the TSKCOR program can be addressed by the CORREG program. Reports generated by three of the new programs can be written on the CODAP REPORT file and later extracted and published using request cards.

Now let's go through a semi-hypothetical example to illustrate how some of the task-oriented CODAP programs might be used in a study. To provide a degree of linkage with the paper being presented by Lt Col Mead, I will describe a study in which the goal is to establish training priorities for police patrolman tasks in the civilian sector. We will assume that supervisory ratings have been collected on three task factors which have relevance for our decision. These are (a) consequences of inadequate performance of each task, (b) task delay tolerance, and (c) task learning difficulty.

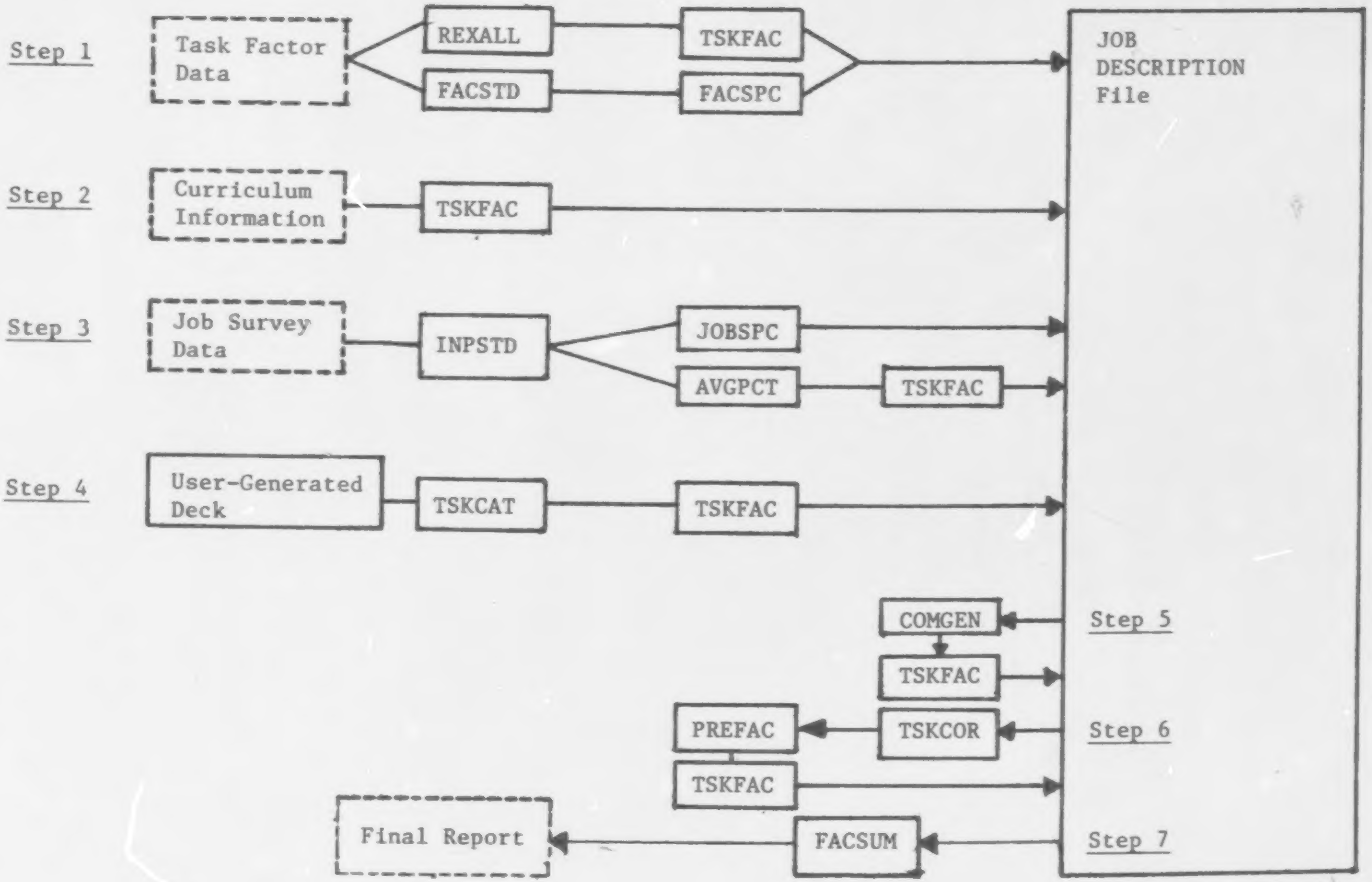
A job inventory containing all of the tasks performed by individuals in the police patrolman career ladder has been administered to a large sample of workers. Instructors in the training academy have indicated those tasks which are currently being taught. Our goal is to recommend for consideration certain modifications in course content and to support these recommendations. Figure 1 presents a simplified schematic of one possible route to the solution. In Step 1, we will process the supervisory task ratings on the consequence, delay tolerance, and difficulty factors. This could be accomplished by going through the REXALL and TSKFAC programs. The mean task ratings for all three factors could be written on the JOB DESCRIPTION file with a single TSKFAC run. An alternative route would be to go through the FACSTD and FACSPC programs, but this would not provide the basis for identifying raters for elimination or computing reliability coefficients. However, the latter route would enable the investigator to compare ratings provided by subgroups. For example, he could compute the mean ratings provided by police lieutenants and sergeants using FACSPC, and then display differences in their perspectives using the FACSUM program. A new program, called REXSPC, is being written to facilitate communication between the FACSTD and REXALL programs. Be that as it may, either of the two routes shown on the schematic can be used to get the mean factor ratings for the total sample onto the JOB DESCRIPTION file.

Since all of the information we have on the course curriculum is a vector of 1's and 0's indicating whether each task is or is not in the present course, we simply punch this information into a card deck and enter it onto the JOB DESCRIPTION file using the TSKFAC program. This is represented in the schematic as Step 2.

In Step 3, we generate two vectors from the job survey data file. First, we compute a job description using the JOBSPC program for patrolmen who have been on the job less than 12 months since graduation from the academy. The "percent performing" vector in this description can be used to estimate the probability that academy graduates will be required to perform each task during their first year on the job, which is one of the most valuable factors for making curriculum decisions.

The second factor generated from the survey data file is designed to identify tasks which are not relevant for consideration for inclusion in the academy course, since they are inappropriate for assignment to privates and corporals. In this case, we apply the AVGPCT program to the grade var[illegible]e. This program computes the sum of the cross-products of [illegible]cent

Figure 1. Schematic of Example Problem Utilizing Task Oriented CODAP Results


Step 1
Task Factor Data
REXALL
FACSTD
TSKFAC
FACSPC
JOB DESCRIPTION File
Step 2
Curriculum Information
TSKFAC
Step 3
Job Survey Data
INPSTD
JOBSPC
AVGPCT
TSKFAC
Step 4
User-Generated Deck
TSKCAT
TSKFAC
Step 5
COMGEN
TSKFAC
Step 6
TSKCOR
PREFAC
TSKFAC
Step 7
FACSUM
Final Report

performing-times-grade values across all grade levels, and divides this total by the sum of percentages. This vector is added to the JOB DESCRIPTION file using the TSKFAC program.

The vector generated in Step 4 using the TSKCAT program is designed to restrict the outputs of the FACSUM displays to be computed in Step 7. Eighteen tasks were selected for this display. They were identified by 1's in an NTASK vector, wherein all other tasks were identified by 0's.

In the sample problem, we shall elect to use the vector of 1's and 0's representing the tasks currently being taught in the academy as the criterion for establishing a training priorities index. This may sound circular, but it is not as naive as it first appears. The current academy course is the end product of many years of curriculum judgments. Our goal is to capture the systematic variance in this judgment process by a regression equation which gives consideration to the factors previously enumerated. We then will apply the equation to every task in the occupational area. By this method, we may be able to identify tasks not being trained which would be trained if the judgment process had been systematic. In the same manner, we may be able to identify tasks being trained which the policy equation suggests should not be trained. The rationale for this approach will be discussed in more detail when we view the final reports generated in Step 7.

It was hypothesized that certain of the task rating factors might be related to the criterion in a non-linear fashion. In order to test this hypothesis, we needed to square certain vectors for introduction into the regression equations. This was accomplished in Step 5 by the COMGEN program.

At this point, we have available on the JOB DESCRIPTION file all variables required for computing the final regression equation in Step 6. This is accomplished using the TSKCOR program. In the same run, the computed regression weights are applied to predictors using the PREFAC program, and this vector is written onto the JOB DESCRIPTION file using the TSKFAC program.

The final results are displayed by reports generated in Step 7 using the FACSUM program. I'm sorry that vugraphs are too restricted in space to permit a full display of FACSUM runs, but I will present information extracted from one run in Tables 6 and 7. First let's look at Table 6.

In the first column, the FACSUM program has listed the

Table 6. Example FACSUM Output

Task Title	Tng Priority	In/Out Course	% Performing	Dly Tolerance	Consequences	Difficulty	Cons(-)Dif	Avg Grade
Conduct Frisk Search	1.14	1	85	1.6	7.3	4.3	3.0	3.5
Subdue Subject Resisting Arrest	1.10	1	69	1.1	7.6	5.2	2.4	3.5
Engage in High-Speed Pursuit Driving	1.09	0	69	1.1	7.0	5.9	1.1	3.5
Fill Out Theft Report Forms	1.00	1	81	5.6	5.5	4.6	.9	3.5
Advise Suspects of Their Legal Rights	1.00	1	68	2.1	6.6	3.6	3.0	3.6
Issue Moving Traffic Citations	.91	1	89	2.4	5.8	3.6	2.2	3.5
Check Autos Against Stolen Car List	.89	1	87	3.2	4.5	2.8	1.7	3.5
Issue Parking Citations	.78	1	51	4.7	3.1	2.8	.3	3.4
Take Statements in Criminal Cases	.59	0	32	4.1	6.3	6.1	.2	3.7
Deliver Departmental Mail	.41	0	29	6.5	2.5	2.6	-.0	3.7
Show Mug Shots to Witnesses	.40	1	16	5.6	5.5	4.2	1.2	4.1
Investigate Repossession Complaints	.30	0	24	6.7	2.5	4.5	-2.0	3.9
Punch Cards for Data Processing System	.22	0	1	6.8	4.5	5.9	-1.4	2.4
Photograph Lineup	.14	0	3	4.6	4.7	4.9	-.1	4.1
Train/Care for Dogs (Canine Unit)	.09	0	2	6.4	4.5	7.0	-2.4	3.2
Conduct Airborne Traffic Surveillance	.08	1	10	3.2	2.6	5.5	-2.9	3.4
Inspect Men at Roll Call	.04	0	3	3.8	4.0	3.5	.4	4.7
Prepare and Present Speeches	-.02	1	7	7.8	3.0	5.9	-2.9	3.9

titles for the eighteen tasks selected for the example using the operator variable generated in Step 4 by the TSKCAT program. Column 2 displays the training priority indexes generated by the TSKCOR and PREFAC programs in Step 6. Notice that all information in Table 6 has been ordered from high to low on the training priority index. The FACSUM program provides the investigator with the option of ordering information from low to high or high to low on any vector--or to list data in duty/task order.

Column 3 identifies tasks currently being taught in the academy by a value of "1," and tasks which are not being taught by a value of "0." This vector was introduced onto the JOB DESCRIPTION file in Step 2 using the TSKFAC program.

Column 4 indicates the percent of patrolmen who have been on the job 12 months or less performing each task. This was generated in Step 3 using the JOBSPC program.

Columns 5, 6, and 7 present the mean task-factor ratings on Delay Tolerance, Consequences of Inadequate Performance, and Learning Difficulty. These mean values were computed in Step 1 using the REXALL program, and were introduced onto the JOB DESCRIPTION file using the TSKFAC program.

Column 8 reports the differences between the values in Columns 6 and 7, and has been included in the table to illustrate one of the many capabilities of the FACSUM program. High positive values in this vector identify tasks which are easy to learn, but for which the consequences of inadequate performance are severe. Examples are "advising suspects of their legal rights" and "conducting frisk searches." High negative values on the vector identify difficult tasks with minor consequences for inadequate performance, examples being "presenting speeches" and "conducting airborne traffic surveillance."

Column 9 presents the average grade values computed by the AVGPCT program in Step 3. It reveals that "Inspect Men at Roll Call" is a sergeant or lieutenant task, and not appropriate for inclusion in the academy entry-level course.

Now let's go back and look at Column 3, which indicates whether each task is or is not included in the present course. Remember that the tasks are ordered in terms of values or the Training Priority Index. Notice that training is already being given on seven of the top eight tasks. The one exception is the third task listed: "Engage in High-Speed Pursuit Driving." No training is presently included in the course on this task. The equation suggests that training on this task should be introduced if curriculum designers are to be systematic in their decision process. The factor data reveal that 69% of graduates encounter this task during the first year after graduation. It has the lowest Task Delay Tolerance among all of the tasks on this chart. This means that when the task is encountered, performance must begin immediately, or the opportunity to perform it rapidly perishes. There is no time to read a manual or consult a friend. The consequences of inadequate performance are severe. I'm not sure of the exact values, but I recall reading in a recent newspaper article that about 50% of high-speed chases end with an accident, and that a significant number end with a fatality. Data on the chart show the task to be difficult to learn, and indicate that it is appropriate for consideration for inclusion in the academy course.

Now let me qualify this statement. The equation actually reflects the need for entry-level training on each task. It does not address the question as to where that training should be given. It is possible that this particular task could be taught more economically in the job setting. Be that as it may, the equation does reflect a need for training.

Looking at the bottom of the chart, we find that training is currently being given on two of the bottom seven tasks. The bottom task, "Prepare and Present Speeches," is performed by only seven percent of academy graduates during their first year on the job. While the task is rated as being difficult to learn, performers are seldom under time pressure (Task Delay Tolerance = 7.8), and supervisors rate the consequences of inadequate performance as being minor. For somewhat the same reasons, the equation raises questions concerning the training being given on "Conduct Airborne Traffic Surveillance," which is performed by only 10% of the graduates.

Table 7 presents an extract of another FACSUM output obtained in the same CODAP run. In this case, a difference score was generated between the observed and predicted training priorities, and the tasks are ordered on that vector. In accordance with user specifications, those tasks with near-equal observed and predicted priorities were suppressed from the printout, leaving only those tasks for which training decisions are required.

I'm sure that you appreciate the impossibility of adequately describing the flexibility and utility of the new task-oriented CODAP programs in the few minutes available to me at this conference.

Table 7. Recommendations for Possible Course Change

Task Title	Tng Priority	In/Out Course	Difference	% Performing	Dly Tolerance	Consequences	Difficulty	Avg Grade
Engage in High-Speed Pursuit Driving	1.090	0	1.090	69	1.1	7.0	5.9	3.5
Take Statements in Criminal Cases	.592	0	.592	32	4.1	6.3	6.1	3.7
Tasks Omitted for Which: Difference: TSKF06(D) - TFAK01(F) is Less Than .500								
Show Mug Shots to Witnesses	.400	1	-.600	16	5.6	5.5	4.2	4.1
Conduct Airborne Traffic Surveillance	.088	1	-.912	10	3.2	2.6	5.5	3.4
Prepare and Present Speeches	-.028	1	-1.028	7	7.8	3.0	5.9	3.9

A UNIVERSAL MODEL FOR EVALUATING BASIC ELECTRONIC COURSES IN TERMS OF FIELD UTILIZATION OF TRAINING

Walter E. Driskill, Thomas J. O'Connor,
Hendrick W. Ruck, and Karl A. Hickerson

Introduction

In support of our primary mission, we train thousands of people. Some of the most expensive training is in the field of electronics. The Air Force alone has about 63,000 personnel involved directly in the electronic fields and the cost of their training runs into the hundreds of millions of dollars. The Air Training Command estimates that Air Force electronic training costs about a half million dollars per day.

Until now, there has not been any reliable method of identifying precisely the amount of electronic theory needed to perform various jobs. As a result, to be on the safe side, we tended to overtrain our personnel. The main purpose of this project is to develop a universal model to evaluate usage of basic electronic principles training.

The criterion used by the model to evaluate electronic theory training is a determination of the usefulness of the training vis-à-vis the performance of assigned tasks in the various electronic career fields.

It is important to be able to identify in a very specific manner those portions of the electronics training which contribute most and least to the performance of tasks. The identification of the relative merit of various portions of electronics training has important implications for managers in the areas of training, personnel classification and testing.

The Electronic Principles Inventory

General Background

The model developed in this project is called the Electronic Principles Inventory (EPI). The EPI or survey is different from the usual task oriented survey in two major respects. First, the EPI asks two general questions: What do you do and what electronic knowledge do you use in performing your job? The usual task survey concentrates on only one question: What do you do? The second difference is that the EPI can be administered to anyone who works with electronics. That is, it is general in nature, unlike the usual survey, which is aimed at a single specialty within a career field.

The EPI is similar to the usual survey in that the data can be analyzed using the Comprehensive Occupational Data Analysis Programs (CODAP).

The EPI contains two sections. Section one has the usual background information, such as rank, command, job title, active military service time, etc. Section two contains the electronic type questions. The content for the questions was taken from the Keesler Air Force Base Basic Electronic Principles Course, 3AQR30020-1. This course is the first military electronics course taken by personnel in 17 specialties within the Air Force and has a documented length of 18 weeks. A few of the specialties skip some of the lessons or modules. The reason the Keesler course content was chosen to be the base for formulating electronic questions was because it appears to be one of the most extensive in terms of the depth and breadth of electronics material.

A typical job description of an electronic specialty may include a general terminology which uses such verbs as monitors, analyzes, identifies, installs, maintains, troubleshoots, repairs, modifies, aligns, inspects, calibrates, isolates, etc. Two questions may be asked. First, does the action verb mean the same thing for different specialties? Second, even if the action verb does mean the same thing in a general sense for different specialties, does each specialty use electronic knowledge on the same depth and breadth dimension? The EPI results are independent of how one would answer the above questions. That is, the EPI asks questions at such a basic level that the data will yield the kind of information one can use to make decisions without being affected by semantic and communication problems that have plagued previous attempts to deal with the complexity of electronics training.

Brief Description of the Model

In general, electronics courses start with the simple and continually build toward the complex. That is, basic components such as resistors, capacitors, etc. are introduced first and eventually they are combined to form such items as power supplies or motors. The following example illustrates how the EPI determines the utilization of electronic principles training.

The example is for motors only but a comparable section has been written for resistors, capacitors, diodes, transistors, etc. The section on motors starts by asking: Does your job involve any tasks dealing with either alternating current or direct current motors? If the individual answers no, he is routed to the next section of the EPI. If the response is yes, the EPI seeks to determine specifically what tasks the individual performs on motors.

The format is as follows:

Do you perform any of the following tasks on motors?

1. inspect . . . . YES NO
2. troubleshoot down to component parts . YES NO
3. troubleshoot as far as checking wire connections but do not troubleshoot down to component parts . . . . . YES NO
4. clean or lubricate . . . . YES NO
5. operate . . . . YES NO
6. remove or replace complete motors . . . YES NO
7. remove or replace motor parts . . . . . YES NO
8. other (specify) . . . . YES NO

____

____

From an examination of the pattern of responses to the above tasks, some obvious assumptions about the level of electronic knowledge needed to do different tasks can be made. That is, an individual who removes or replaces complete motors generally doesn't need to know as much about motors as an individual who troubleshoots down to component parts.

The EPI also seeks to determine if the individual performs tasks on specific motor parts such as the field coil, armature, rotor, brushes, slip rings, commutator or pole pieces. Subsequent questions ask whether the individual is involved in determining magnitudes and directions of torque and induced voltages in motors. The final set of questions on motors asks about the types of motors the individual works on (synchronous motor, induction motor, etc.).

The complete data on motors and all the other sections are analyzed to determine field utilization of the electronic principles training.

Administration of the Electronic Principles Inventory (EPI)

Initial Field Test

In order to assess the universality of the instrument, it was decided to administer the inventory to personnel in career fields requiring various amounts of electronic knowledge. Figure 1 gives the specialties in the communications-electronics systems, missile electronics maintenance, and avionics systems career fields which were sampled.

307X0	Telecommunications System Control
316X1L	Missile System Maintenance
324X0	Precision Measuring Equipment (PMEL)
326X0	Avionics Aerospace Ground Equipment (AGE)
328X3	Electronic Warfare System

Figure 1. The electronic specialties which were field tested.

Another variable addressed in selecting the sample for the initial field tests was Major Air Command. Airmen serving in the Strategic Air Command, Tactical Air Command, and Aerospace Defense Command were surveyed. These three commands represent the bulk of USAF airpower in the continental United States.

The EPI was administered to small groups of airmen (N=200) by project members at seven different locations throughout the United States. Average time needed to take the EPI was one and a half hours.

Full-scale Administration

Once the EPI has been reviewed and finalized it will be administered worldwide to all Electronic Warfare System personnel (AFSC 328X3) and all Telecommunications Control Systems personnel (AFSC 307X0) in the Air Force. Preliminary data from the initial field test indicate that significant results regarding training and personnel classification can be found for these specialties. It is expected that the worldwide administration will take place in December, with initial data being available in March 1976.

Results and Applications

Since the results presented here are based on a limited field study, the emphasis in this section will be on potential applications of the results of surveys using the Electronic Principles Inventory (EPI).

Electronic Principles data can impact meaningfully on several major functions which include technical training, personnel classification, and promotion testing.

Technical Training and Personnel Classification

Preliminary results indicate that personnel in different specialties use different amounts of electronic principles theory in performing their jobs. Figure 2 shows the average use of electronic principles theory, which is indicative of the probability of use by any one individual within a specialty. The range of average use is from 10% to 51%.

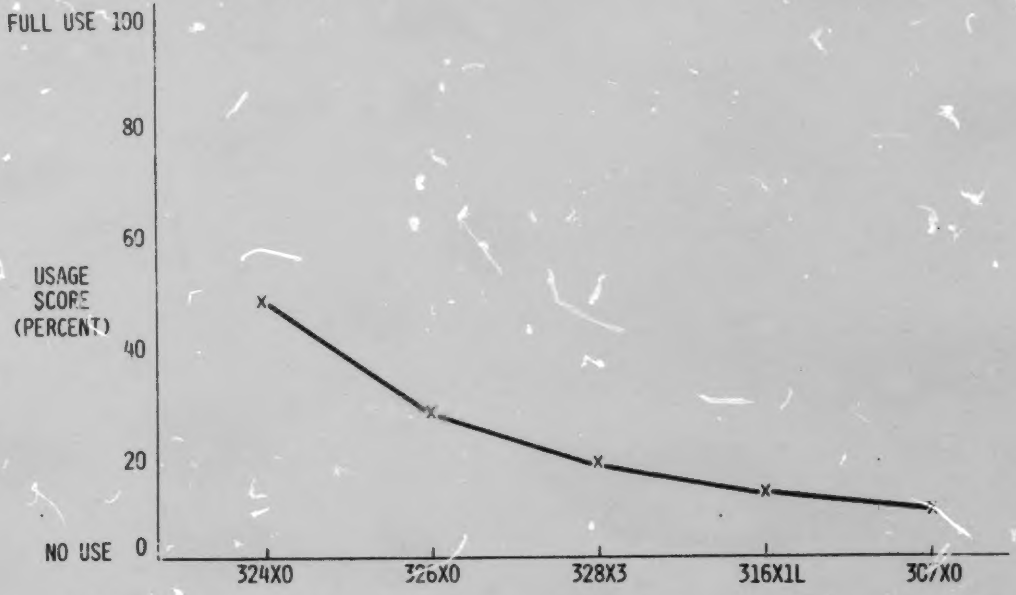


Figure 2. Average use of electronic theory by five specialties in the Air Force.

Training organizations, however, are more interested in the total use of electronic principles theory by all individuals in the specialty. The data shown in Figure 3 indicate a range of total usage between 14% and 84%.

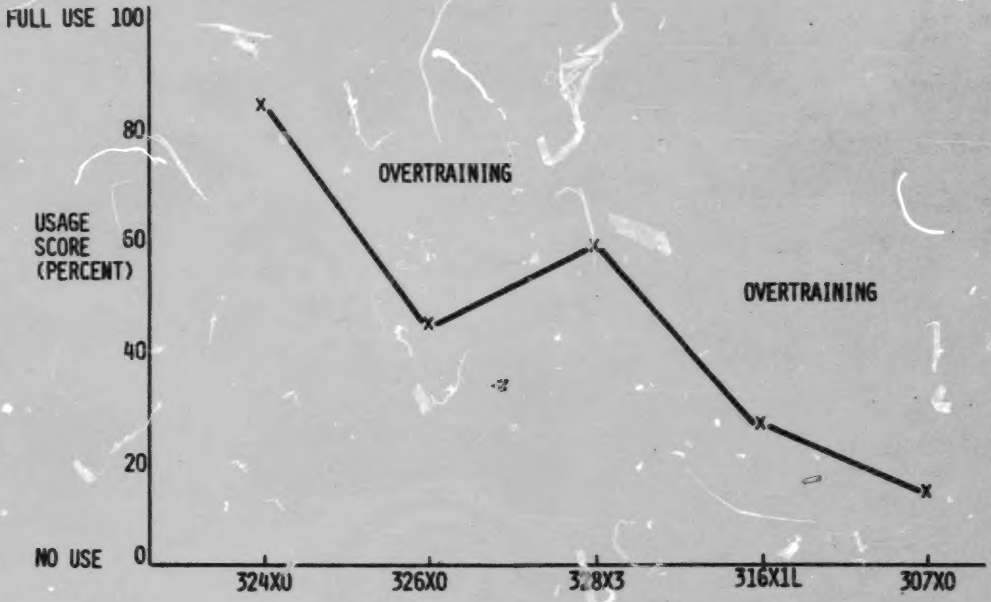


Figure 3. Total use of electronic theory by all individuals in five Air Force specialties.

The percent or area above the curve represents overtraining. That is, nobody in specialty 324X0 indicated a use for 16% of the electronic principles theory; 86% of the taught electronic principles theory is not used by specialty 307X0 personnel. Relating the data on the curve to the mental framework for analysis developed in Appendix A, one can see that the area below the curve represents the depth and breadth of electronic principles knowledge used by the specialties.

In support of the conclusion that nobody uses the theory above the curve, the specific information (16%) that specialty 324X0 does not use was identified. The group of 324XO specialists that writes the Specialty Knowledge Test (SKT) was asked questions about this information. They indicated no use of, familiarity with, or relevance for such information in their job. Yet, one can go to some of the basic electronic courses and find that an individual cannot progress through the course unless he demonstrates his comprehension of such knowledge. Referring to the model developed in this project, it can be shown that much of this extraneous course information is presented at a depth below the level of practical use, that is, at the structure and process MICRO level (in Appendix A), which generally is of importance only to design engineers.

To further validate the conclusion that information above the curve represents overtraining, the extraneous information was compared with the course material of an electrical engineering department at the college level. The comparison itself was made by the faculty of the college. They concluded that: "Only about 5% of the extraneous information is taught here [at the institution]; the rest is physics theory."

The data also indicate that most of the knowledge is cumulative between specialties. Take, for example, the knowledge used on the job by 307XO personnel and compare it with the knowledge used by 324XO. The 324XO personnel use the same 14% of the knowledge which is used by the 307XO personnel, and an additional 70%.

Figure 4 combines the two previous curves and gives an overview of the full impact of the data.

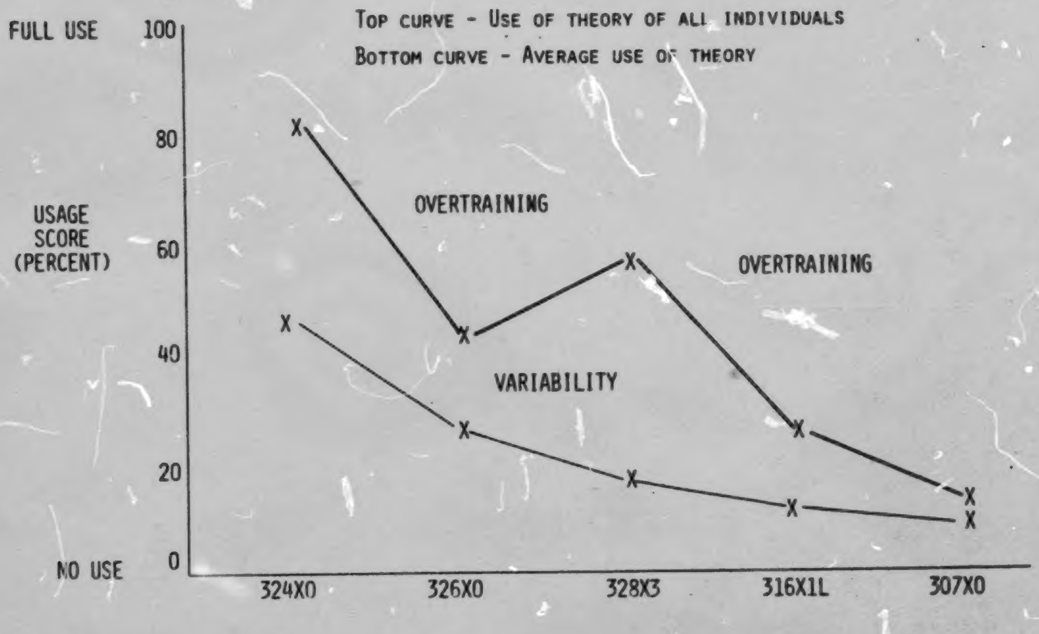


Figure 4. Total use and average use of electronic theory by five specialties in the Air Force.

The data can be divided into the following categories:

1. The percent or area above the top curve represents overtraining within each specialty.

2. The percent or area below the top curve represents the total usage made of electronic theory training within each specialty.

3. The percent or area between the curves represents the degree of variability of use of electronic training within each specialty. Variability for each specialty is determined by subtracting the appropriate value on the bottom curve from the value on the top curve. This value gives a difference or variability score whose utilization will be discussed shortly.

4. The percent or area below the bottom curve represents the average use made of electronic theory training within each specialty.

Table 1 presents the same data as the curves in Figure 4. One obvious use for this information is in formulating a hierarchy of specialties based on technical difficulty. Since there is a large amount of overlapping of the uses of electronics theory between specialties, the establishment of a hierarchy of specialties based on technical difficulty can provide managers with data to make decisions on determining the most equitable moves (forced or volunteer) between specialties. For example, when one specialty needs to be filled, the question arises: What other specialties are the closest in terms of use of electronic knowledge? Specifically, if vacancies in 326X0 existed, it would be more logical to move people into that specialty from 324X0 and 328X3 than from 307X0. If it became necessary to fill 326X0 slots with 307X0 personnel, a refresher basic electronic course would seem desirable. On the other hand, if vacancies existed in 307X0, any of the specialties listed above 307X0 could easily fill the slots after an equipment orientation.

Another use for the data is in the determination of whether to combine or shredout various specialties. For example, a specialty with a high usage score and a high difference score is indicative of a specialty which uses a high level of electronic theory, and in which groups of individuals within the specialty use different portions of electronic theory. Based on this data, one would perhaps recommend a shredout of the specialty. The data can also be used to give managers an indication of the amount of cross-training needed by personnel within a given specialty. A low variability or difference score indicates that personnel in that specialty can be moved around easily. For example, for 307X0, the degree of variability within the specialty expressed as a difference score is 3.23. Compare this with 33.31,

Table 1

Percent of Average Use and Total Use of Electronic Principles Training by Five Air Force Specialties

AFSC[a]	Average use	Total use	Difference
324X0	50.65	83.96	33.31
326X0	31.18	43.98	12.80
328X3	20.18	57.68	37.50
316X1L	15.09	25.80	10.71
307X0	10.49	13.72	3.23

Note. According to Air Force Manual 39-1, all specialties listed require the same minimum electronics score (80) for entry into the specialty.

[a] Air Force Specialty Code

the difference score for 324X0. What this comparison means is that, based on the use of electronic knowledge, personnel in 307X0 are comparatively easily cross-trained within the specialty, whereas, personnel in 324X0 are not. This conclusion has been partially verified by interviewing 307X0 and 324X0 specialists.

This data could also be used as inputs to either of the following decision processes: (1) that which involves a reduction-in-force (RIF); (2) that which involves who will be allowed to reenlist (selective reenlistment). Other things being equal, it would make more sense to retain those people with a higher experience level in electronics. High usage scores are indicative of a career field with a high experience level in electronics.

Another application for the data shown in Table 1 relates to the initial placement of personnel in various specialties. Presently, all the listed specialties require the same minimum electronics score (80) for entry into the specialty. The hierarchy of specialties based on electronic difficulty gives program managers options such that they could place persons with higher electronic aptitudes in the more difficult electronic specialties. For example, the 324X0 (Precision Measuring Equipment) specialty could receive a larger proportion of persons with higher electronic scores. Since the 324X0 personnel perform the critical calibration tasks for most of the electronic equipment used by all Air Force specialties, the concept appears sound.

Impact on Interservice Electronic Training

The model could be used to initiate a feasibility study for determining the extent to which electronic specialties in each service overlap in their use of electronic theory. Based on this information, decisions could be made concerning joint training. If joint training is implemented, the model could be used to evaluate continually and update the training programs. For those military specialties whose training it is not advisable to combine, the model can, as a minimum, provide information on those portions of electronic theory which are most used, least used, and not used.

Specific Comments on Training

In general, electronic principles training is broken into a series of discrete modules, sections, or lesson plans. A trainee for a specific Air Force specialty may be required to pass any subset of these modules before being allowed to train on the equipment he will work with in the field. The EPI has been constructed in such a manner that its sections correspond with the modules. Thus, the data are immediately usable by the training units. Table 2 illustrates how one could determine which lessons or modules of training are required by field personnel to perform their jobs. The example is for AFSC 328X3 (Electronic Warfare Systems personnel); the last column of numbers shows the percent of each module which is actually used in the field by all 328X3. Further, the data can be broken down by command, as shown by the first and second columns, which compare Strategic Air Command (SAC) 328X3 personnel with Tactical Air Command (TAC) 328X3 personnel. The data could be broken down by any chosen variable, such as time in service, grade, etc.

Table 2

Percent of Theory Used by SAC and TAC

Electronic Warfare Systems Personnel (328X3)

Subject	SAC 328X3	TAC 328X3	SAC AND TAC 328X3
D. C. and Voltage	44	44	44
Multimeter Uses	60	60	60
Meter Movements	75	75	75
Transformers	37	47	50
Relays	50	50	50
Microphones	0	40	40
Speakers	0	44	44
Motors	63	0	63
Generators	23	0	23

One can also compare usage of modules between specialties. Both specialties, 328X3 and 307X0 (Telecommunications System Control) take the same course with minor variations. The first column of numbers in Table 3 shows the percent of theory needed by 328X3; the second column shows the percent of theory needed by 307X0; the third column shows the combined need for theory by 307X0 and 328X3. The third column of numbers gives an indication of the overlap of theory, and thus is useful in making decisions about combining basic electronic courses. The data indicate that the 307X0 personnel are not using any of the listed theory portions of their basic electronic course beyond speakers. Although this example is from limited data collection, the results are corroborated from the occupational survey done on AFSC 307X0 (AFPT 90-307-081) which shows an extremely limited basic technical role for 307X0 personnel. In addition to the obvious conclusion that 307X0 personnel are overtrained, other inferences can be made.

Table 3

Percent of Theory Used by 328X3 and 307X0 Personnel

Subject	328X3	307X0	307X0 and 328X3
D. C. and Voltage	44	33	44
Multimeter Uses	60	60	60
Meter Movements	75	83	83
Transformers	50	16	53
Relays	50	0	50
Microphones	40	13	40
Speakers	44	31	50
Motors	63	0	63
Generators	23	0	23
Diodes	45	0	45
Transistors	59	0	59
Transistor Amplifiers	81	0	81
Electron Tubes	38	0	38
Power Supplies	78	0	78
Saturable Reactors and Magnetic Amplifiers	6	0	6
Synchro-Servo Systems	0	0	0
Transmission Lines	56	0	56

Table 3 (Continued)

Subject	328X3	307X0	307X0 and 328X3
Antennas	65	0	65
AM Systems	91	0	91
FM Systems	91	0	91
Single Sideband Systems	90	0	90

First, if the present course for 307X0 personnel were reduced in scope, it is probable that fewer students would be eliminated due to academics. Second, shorter courses mean quicker utilization of personnel in the field and reduced training costs. Third, persons, such as those taking the longer course required for 324X0 and who are eliminated for academic reasons, could be considered for the 307X0 specialty, or other less demanding (in the electronic sense) specialties. Thus, the time and money already invested in the person could be salvaged.

At a more refined level of data reduction, the EPI gives a hierarchy of percent responding yes to specific EPI statements. Figure 5 shows those items in the EPI to which 50% to 60% of the 328X3 sample responded yes. The items shown reflect the actual field use of the material taught, or that which should be taught, in the basic electronic course.

On the other hand, Figure 6 shows those items which are taught in the basic course and which nobody in the field uses.

At present, we have little or no information about the utilization of basic electronic theory for individuals or groups over time within any given specialty. The curves shown in Figure 7 demonstrate another application of the EPI data. The curves represent hypothetical usage

DO YOU WORK WITH DIODES ON YOUR PRESENT JOB?

DO YOU REFER TO OR NEED TO HAVE A KNOWLEDGE OF:

- HOW TO REPLACE DIODES ON A CIRCUIT BOARD?
- HOW TO REPLACE DIODES ON A CHASSIS?

DO YOU REFER TO OR USE:

- TRANSISTOR SCHEMATIC SYMBOLS?
- TRANSISTOR SUBSTITUTION INFORMATION?

DO YOU WORK WITH TRANSISTOR AMPLIFIERS ON YOUR PRESENT JOB?

FIGURE 5. REPRESENTATIVE ITEMS TO WHICH 50% TO 60% OF THE ELECTRONIC WARFARE SYSTEMS PERSONNEL RESPONDED YES.

DO YOU REFER TO:

- NUMBER OF ELECTRONS IN A PARTICULAR SHELL OR ORBIT?
- ATOMIC NUMBER?
- CONDUCTION BAND IN SEMICONDUCTOR MATERIALS?
- ACCEPTOR IMPURITY IN SEMICONDUCTORS?

DO YOU USE OR REFER TO THE TRANSISTOR GAIN GAMMA?

DO YOU CALCULATE THE FOLLOWING TRANSISTOR GAINS:

- ALPHA?
- GAMMA?

DO YOU COMPUTE THE STATIC OPERATING POINT (Q) OF A TRANSISTOR AT DIFFERENT TEMPERATURES?

DO YOU USE OR REFER TO THE ELECTRON TUBE PARAMETER CALLED AC PLATE RESISTANCE?

DO YOU CALCULATE AC PLATE RESISTANCE?

FIGURE 6. REPRESENTATIVE ITEMS IN THE EPI TO WHICH NOBODY IN THE SAMPLE OF 328X3 RESPONDED YES.

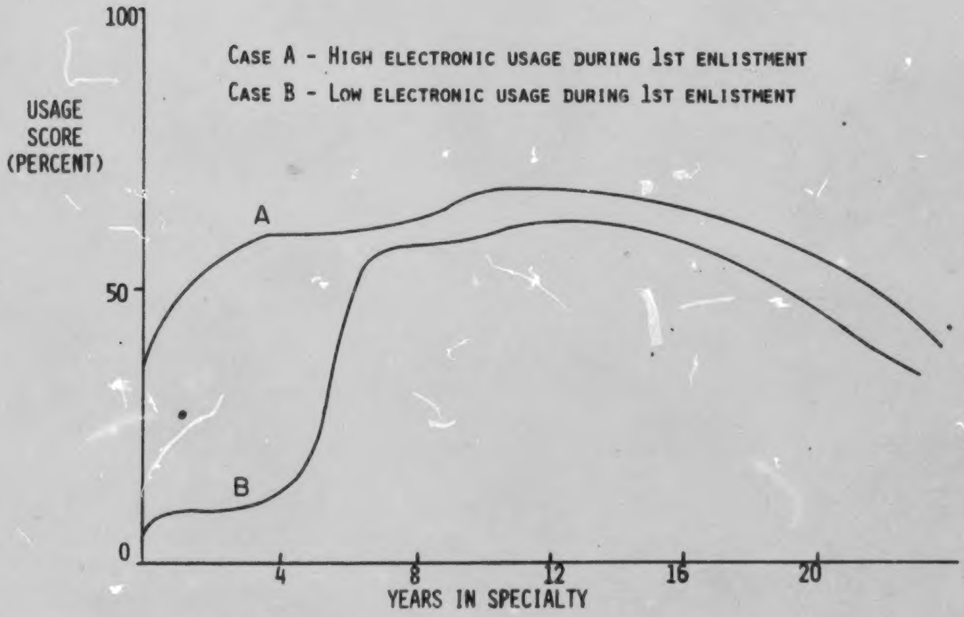


Figure 7. Two hypothetical usage curves for a given specialty over a 20 year period.

for two possible cases. For Case A, use of electronic theory is high in the first four years; thus implying the need for a rather extensive basic electronic course immediately after enlistment. Case B illustrates a low initial usage for the first few years, with accelerated usage after the four year point. Since many military personnel depart the service after their initial enlistment, it might be worthwhile, for Case B, to delay extensive theory training until after the individual reenlists.

Related to training, another use for the EPI data is in the determination of what should be taught in a new course. In the hypothetical case of a new electronics specialty, a manager could have the writers of the technical and maintenance manuals fill out the EPI. The resulting data could be used to form the core for the new course.

The data from the EPI can also be used as a basis for interaction between the training managers and the using agencies. It is not uncommon for the using agencies to request that the training people reevaluate their basic electronic courses. Such requests may state: "Give your students more theory" or conversely, "Give your students less theory." Specifically what is meant by such general words as theory, basics, principles or fundamentals is not usually clearly defined. Even less clear is the task of identifying which portion of the theory course is to be taught more, or which portion is to be taught less. The model developed in this project, when applied by the training managers, is able to identify in a very specific manner the most field-relevant and the least field-relevant information in the training course. Likewise, the results from the EPI will enable the using agencies to levy more realistic training requirements on the training programs.

Promotion Testing

Specialty Knowledge Tests (SKTs) are used in the Air Force as one factor in the airmen promotion system. These tests traditionally are composed of basic electronic theory questions as well as specific equipment questions for each of the electronic specialties. At present, there is no hard data on which portions of the basic electronic theory apply most to a given specialty. EPI data can provide guidelines for establishing specific sets of knowledge items which may be used for knowledge tests for given specialties, and thus support or increase present test validity. Within the same specialty, the EPI is able to determine which basic electronic questions are the most equitable to ask. For example, refer back to Table 2, which compares SAC and TAC electronic warfare specialists (328X3). From Table 2 it can be seen that SAC 328X3 personnel do not work with microphones and speakers, while TAC personnel do not work with motors and generators. Each group would be at a disadvantage if asked questions about items they do not work with. In addition to the selection of the most equitable test items for a specialty, the EPI also gives the testing people an indication of the risks involved in the testing procedure by not having certain commands represented when the tests are being constructed.

One of the implications of establishing sets of theory questions which are job relevant for specialties would be to have a panel of electronic theory experts develop an item bank of electronic theory test items for all electronics specialties. Specialists could then select items from the item bank for their own specialty, thus reducing their workload, since they would only have to develop equipment test items. The bank would be of such a size that any possible compromises of a test could be handled by switching to a new set of basic electronic principles questions.

Since the SKTs are made by groups of personnel on temporary duty status, a reduced workload can be interpreted in terms of money saved.

Conclusion

The Electronic Principles Inventory (EPI) represents a major contribution in the identification of the type of electronic training needed to perform any job. It is general in nature, and thus can be administered to anyone who works with electronics, regardless of the level of involvement (maintenance, operations, training, etc.).

The data from the EPI can be used:

1. By training organizations to evaluate existing programs and to make decisions about the organization of new programs.

2. By managers interested in the feasibility of combining training programs.

3. By field organizations to evaluate the utilization of training and to determine what is needed in terms of deficiencies.

4. By people who work on personnel classification systems to determine optimum utilization of manpower between and within specialties.

5. By testing personnel to determine which portions of the electronic theory relate to actual utilization.

NAVAL MANPOWER UTILISATION UNIT
CREASY BLOCK SOUTH HMS VERNON PORTSMOUTH
PO1 3ER

THE USE OF OCCUPATIONAL ANALYSIS OUTPUTS TO ASSIST TRAINING DESIGN IN THE ROYAL NAVY

Commander C. J. Nicholl - Royal Navy

Abstract

This paper describes how we, in the RN have coordinated the processes of job analysis with training design during the past three years and are beginning to extend the use of these existing techniques with the use of CODAP for the Job Survey and the incorporation of selected sub-routines into some of the training analysis and design stages. The Naval Manpower Utilisation Unit, which I represent, is responsible for the development of job survey and analysis techniques. The Royal Naval School of Educational and Training Technology is responsible for the development of training analysis and design techniques. This paper therefore results from the accumulated experience of both units now working together to ensure that the job analyst produces data suitable for training design and the training designer knows how to make the best use of the data when it arrives.

Over the past few decades, there have been several factors which have steadily gained in significance to the design of training.

The complexity of systems has increased, which means that the determination of the correct content of training courses involves a detailed examination of the individual tasks comprising the job so as to ensure that all necessary elements are covered and that numerous small items of non essential knowledge do not accumulate into an unacceptable amount of over training.

The variety of different tasks undertaken in the range of complement billets by a given specialisation and rate has increased. Thus instructional staff can no longer be expected to have personal experience of the full range of tasks for which they are conducting training. To ensure that a course contains a balanced content, the instructors must have had close guidance based on a breadth of experience.

Job requirements change more rapidly than in the past. This means that a significant element of course content must be predictive in origin while other elements will soon become obsolete. A leisurely process of course design could result in an out of date and possibly misleading course.

Resources are becoming ever more strained as equipment and ships escalate in cost and the changing pattern of society tends to shorten the time which men are prepared to serve in an employment which demands separation from their families. The proportion of the Navy's resources tied up in training is high so improvements in training efficiency can contribute significantly to the efficiency of the operational Fleet.

Job requirement and the efficiency of training can only be effectively controlled if the performance of ex-trainees in the Fleet is constantly monitored in order to update and improve training.

It was in recognition of these factors that the Royal Navy introduced in 1971 objective training, a necessary preliminary being some form of job analysis. This can take several forms but in the Armed Forces, especially when dealing with personnel where levels of skill are wide ranging, a form of questionnaire based survey is an effective tool which provides a reasonably satisfactory answer in terms of the data produced and it is also a technique which designers can handle within the project time scale. However it must be recognized that a job analysis currently takes up 30 to 50% of the training design effort where man hours spent may be up to 20 times the length of course produced. It must also be recognised that the job analysis is record of peoples' perception of their own experiences. However accurate the method, the results will be based on a peacetime situation and may not necessarily be the best source of information for training design. If we are simulating war there is a danger of compounding the errors of the simulation. Further the job analysis results must be modified to take account of emergency tasks or the introduction of new equipment. It must be a matter of judgement whether a questionnaire based survey is the best source of information for training design.

The aim of naval training is to train men to meet the employment requirements of the Fleet, taking into account other factors such as satisfactory career structures as well as resettlement and recruiting requirements which may not be directly specific to the job. Training design contributes to the aim by:

Designing shore courses and where necessary on job training to bring men to the standard approved by the Fleet.

Providing objective, that is precise, measureable and unambiguous statements suitable for:

. The approval of training courses and on job training.

The control of the conduct of courses and on job training.

The assessment by employers of the capabilities and limitations of men.

The monitoring by trainers of the efficiency of their courses and of the achievements of the trainees both under training and in their jobs and also the monitoring of changes in the Fleet's requirements.

Before designing how training designers set about their work it is necessary to describe in some detail the outputs from a job analysis dedicated to training design. For a survey where both population and job boundaries are relatively small there are several custom built job analysis programs which list either time or frequency involvement task by task and which provide the training analyst with adequate quantified detail for his work. However for surveys of large numbers or complex organisations the Comprehensive Data Analysis Program (CODAP) has many advantages. Not only are there many computing routines in the package, but it also has the particular advantage of holding data on magnetic tape available for specific user demands. Only with the aid of CODAP can a complete Branch of the Naval service be analysed making it practicable to conduct job analysis of the entire Fleet on a rolling programme within a reasonable time scale.

It would be inappropriate and probably unnecessary to describe CODAP in detail to this conference but this presentation would be incomplete without giving an outline of the process and of some of the computing routines.

The computer analysis depends on three types of input data; first task time responses which are assessments by the man of the relative time spent on each task with which he is involved. Secondly "secondary factor" data which can be any factor chosen by the analyst to amplify task involvement data such as level of involvement (supervise, do or assist) or the perception of difficulty or importance of each task in which a man is involved. Thirdly, responses to "variable" questions varying from service and personal background, through job aids and attitudes to opinions on Service Conditions or Job Satisfaction. The latter type of input data and the associated computer processing seem to be more widely used by the Royal Navy than other CODAP users.

Before making any analysis of the data, CODAP first needs to establish the Input Standard. The computer is primed with the questions and response options and the raw data from completed questionnaires. This is sorted and catalogued case by case in preparation for subsequent analysis. A printout is provided to give sufficient information to identify cases which are in the wrong format or which may need to be removed from the file for some other reason. The training analyst needs to be involved with the Input Standard procedure because he will be the customer for the eventual printouts, and the grouping of tasks is determined at this stage. The training analyst may want tasks grouped by equipment or by techniques or conceivably by importance or perhaps difficulty. It is always practicable to reorganise the file but it is time consuming both for the analysis staff and the computer so it's better to get the most useful sort the first time.

The fundamental printout for the training analyst is the Job Description. This lists for selected populations the percentage who carry out each task and converts the men's assessments of relative time involvement (very little to very much) to a percentage time involvement, both for those involved and for the whole population. The task can be ordered in several ways. The preamble includes the sum of task and time involvement in each group of tasks. The CODAP Job Description is similar to that which other job analysis programs provide and has the same disadvantages that there is an enormous amount of data which it may be necessary to consult in detail at some stage, but which demands time-consuming laborious inspection by the training analyst. It is our experience that time-constraints or perhaps lack of enthusiasm have on occasion caused the training school to put such job analysis information to one side and then get on with subjective course design based on experience. However once populations have been defined by Job Description CODAP provides a number of useful extracts:

Group Summaries where task or time involvement of up to 14 populations are listed side by side to assist in comparison.

Group Differences which compare two selected populations and print the difference between the time or task involvement in descending sequence. Tasks with similar involvement need not be printed.

Diagram which is the printout of a clustering process which has grouped up to 2000 individuals according to the simularity of the work content. The printout is a tree diagram which is in effect a model of the branch under study. Each group in the model can be treated as a population to call up a job description or other relevant sub routine.

History variables - that is data on mens background, usage of tools or attitudes are analysed and displayed as a Variable Summary which compares the responses of up to 12 populations giving the raw scores for each response option, the mean of responses and standard deviation.

Secondary factors are analysed and displayed with detail similar to history variables.

Having described just six of the twenty five computing routines in CODAP I will lead into how they can be used in the process of training design. This is not to say that the others are not suitable, but it is unlikely that the design teams would call for the more esoteric routines until they are reasonably experienced in the use of CODAP as a tool.

The first stage of training design must be involvement in the questionnaire construction. The make up of the inventory of tasks and the response options will be fundamentally effected by the decisions which are to be made on the basis of output from the data bank and the necessity for the task inventory to cover the whole job and to be written on the correct level. The training designer will often wish to go to such a fine level of detail that the questionnaire would be unacceptably long. On the other hand a shorter questionnaire with generalised task descriptions could be useless for training design. A balance must be struck.

The response options against each task need careful thought. Perhaps a man should only mark what task he does and at what level of involvement. Difficulty, importance or distastes are perceptions affected by the situation of the time, earlier training methods and individual characteristics. On the other hand if a task is considered by a substantial proportion of a sample to be difficult and rather nasty perhaps it is. A problem of using several response options as secondary factors for each task is that each option is making a substantial increase to the time which man will take to complete the questionnaire - perhaps by as much as 50%. It may therefore be better to group time rated tasks together and seek responses on level of supervision, difficulty or perceived importance as history variables.

At the questionnaire construction stage it will also be necessary to decide whether amplifying information will be collected from supervisors or employers. A criticism of job analysis is sometimes voiced that it is trying to promote attempts to run the Navy on the evidence of responses from a lot of ignorant sailors and inadequate notice is taken of the experience and wisdom senior ratings and officers. Of course those with a width of responsibility can best judge the true importance or criticality of a task and whether it should or should not be done - but it is often revealing to find that the shop floor perceives the all important to be mundane and the trivial to be vital.

Having helped with developing the questionnaire, the training designer can retire to the wings while the analysis unit gets on with administration to the sample, the transfer of data to the computer and the successful achievement of the Input Standard routine.

Except for straightforward redesign of existing courses the job structure should always be examined. The clustering program Diagram will show groups which have a common work content. If there is an anomaly between this and the established structure the solution could be to reallocate tasks between rate or specalisation taking work content and time into account. Appropriate tasks could be identified from Group Summaries or Group Difference printouts. Another solution could be to recategorise either the specialisation or rating allocated to complement billets using data from Job Description to quantify the implication of any change.

Components other than task time involvement data should also be examined including the comparison of usage of tools or techniques and levels of supervision. The Variable summary of history Data and Analysis of Secondary Factors will be useful.

If importance or difficulty of task have been assessed independently by supervisors or trainers the level can be marked against a Group Summary to check that they are appropriate to the structure. It would be wrong for a junior rating to spend more time on tasks of high importance than his seniors and it would similarly be sensible to give the more difficult tasks to the more experienced man, regardless of time involvement. On the other hand, if a job description is restricted to mundane tasks, however junior the rating, disenchantment will set in.

If the questionnaire includes responses from ratings on the job satisfaction it might be practicable where one group is showing less satisfaction than others to adjust the job structure.

Having reached on agreed job structure the Operational Performance Standard (OPS) can now be developed. The following steps would be followed:

Identify non professional tasks in the inventory. While in general you would not expect to include these in the OPS there would be some which are shown to be significant in the Job Description and which might be appropriate for training.

Identify tasks from the Job Descriptions or Group Summary where there is low time involvement. I beg to differ from the USAF where, as I understand it, the training given directly reflects the percentage of people involved in the task. Perhaps this is an over simplification. If there is low time involvement in an important task, it may require special emphasis in training to help the retention of memory. As an example few sailors (hopefully) actually fight fires, but all must be given thorough practical training.

Adding emergency, war or new tasks including any task to be encountered before the men concerned are due for retaining.

Group together tasks into an appropriate level of detail for the OPS.

Assess the conditions under and standards to which the tasks need to be performed. There may be some guidance from job analysis data.

Note from Group Summaries any parts of the OPS which apply to specific ship types.

Having sought and been given approval for the Operational Performance Standard, the training designers next stage is to allocate tasks to Career, pre-joining or on-job training. The Group Summary printouts will be a valuable aid to decide whether tasks are specific to certain ships or equipments and so suited for pre-joining on-job training rather than career. The choice must be made between sea (on-job) training or shore (pre-joining) training. A bias towards shore training would go to tasks which are:

Important. Importance cannot be found directly from the job analysis but the trainees experience can be enhanced by the results of a limited survey to supervisors or employers. In a perfect world only the Fleet Commander or his Staff can grade the importance of tasks.

Low frequency of performance. A task which is frequently performed and where the consequences of error are not critical can be learned on board. It is likely to come up and be learned during the turn over or work up period. The frequency of performance can sometimes be inferred from the task time involvement, but if not, it may be necessary to include specific questions on frequency in the questionnaire. We have found a need for this in the usage of tools or techniques such as precision lathe work, reading simple wave forms on a oscilloscope or interpreting wiring diagrams. These would form a part of many of the tasks in the inventory and so cannot be included as specific tasks in themselves.

Unsupervised tasks which will be apparent from the analysis of secondary factors on level of involvement.

Tasks which are difficult to learn or difficult to perform but depending of course on the training facilities available.

The training designer would be biased to allocate to on-job training tasks which come at the other end of the scale of importance, frequency, supervision or difficulty. However each task must be taken on its merits; we have tried at the NMUU to write a training designer's Algorithm to help the decision process which seems simple in the abstract but which in practice we found too complicated for an organisation which has no research function.

To summarise, the training designer would be predisposed to allocate tasks as follows:

Tasks with wide application to Career training.

Tasks with limited application to Pre-joining or On-Job training.

Important, Unsupervised and Difficult tasks to Shore training.

Tasks where the consequence of error is slight, tasks where the man is assisting or which are repetitive to On-Job Training.

Finally the training designer identifies from Group Summaries or Group Difference outputs the tasks which have already been learned in practice, and is ready now to design the actual course giving training emphasis to the same patterns as the decision whether to train ashore or on the job, using the same job analysis outputs. He should also take into account attitudes where they are significant.

I will round off this presentation by describing some examples of how we have used job analysis for training design in the Royal Navy.

The first example is from the survey of the job of the Supply Officer. The first stage was to study the Junior Supply Officer who has previously given across-the-board training in secretarial duties, accountancy, storekeeping and catering. The analysis confirmed that the majority of these officers, who are appointed to secretarial billets, have little or no involvement in the other aspects of their specialisation. The result of the survey has been to redesign the training on a modular basis, all being given secretarial training in detail, but teaching on other aspects is now limited by excluding the particulars which would only be forgotten without actual involvement. The few officers appointed to accountancy, storekeeping or catering billets are now given a special job module of training which can be in greater detail than before.

The next stage of this analysis was to look at the more senior officer appointed in Supply Charge at sea. Here the employment is very different. There is a very broad spectrum of tasks in which they are involved and the Supply Charge Course must cover them all, although lesser emphasis can be given to secretarial duties, not only because the officers would have had previous experience as a Secretary but also because there will always be another officer in the ship carrying out the duties of Captains Secretary. The job analysis while confirming what was generally recognised will provide a good basis for redesign of the Supply Charge Course. However from examination of the results of this study it seems to me that the analysis of jobs higher in the Management structure which have a progressively lower "hands on" content, which by their nature will demand a differing emphasis on tasks according to the situation - and by this I do not mean management by crisis - and which will be held by progressively fewer men as they get more senior, each with his own style of management and each with a greater weight effort on the sample, requires a different method which may or may not include a questionnaire. Even with a questionnaire the behavioural style of task inventory which is the most useful for the analysis of operator jobs would not form a major part of the middle or upper management job analysis.

My next example concerns on one part, the ancient trade of the Naval Shipwright traditionally responsible for the woodwork and metal work of the ship's hull and boats, indeed they are now known as Marine Engineer Artificer (Hull), and on the other the not quite so ancient trade of the Marine Engineer Artificer (Propulsion) responsible for keeping the main engines and the many bits of auxiliary machinery going. Two or three years ago it was decided that the MEA(H) job could be equally well done by the MEA(P) with little extra training and so a common MEA was introduced. Last November we conducted a job analysis of the Engineering Branches using CODAP and from this have been able to compare the work of the MEA(P) and the MEA(H) in quantified terms. This has shown that while the MEA(P) is involved in only a few tasks in each of the several work areas associated with his trade, the MEA(H) is involved in many tasks in only a few areas. The jobs are different and the training should be different. The examination of the work of the composite MEA now going to sea has shown that he is being employed as a machinery watchkeeper in order to gain his necessary certificates which are essential for his promotion and so is inevitably foregetting the skills of metalwork, welding and joinery taught in his training. There are thus both branch structure and training implications from the survey.

My final example concerns the gathering of attitude responses and the value of job satisfaction data to training designers. Due to a temporary surfeit of recruits in certain categories there were for a time insufficient billets at sea for all the products of the training machine. Some were being held ashore for a time which was recognised as a regrettable situation. The Engineering Survey included 37 questions on Job Satisfaction and responses were compared between those young men held back ashore with those more fortunate and drafted to sea. It was not surprising that there was a marked chill in the climate of satisfaction for "the work you do now is what you expected on completion of training" "The importance of your present job" "The fairness of Naval Drafting" and "Your present life and work in the RN". It was also to be expected that there would be a highish mark for "The adequacy of training given for your present job" - (anyone can sweep leaves). What was unfortunate was a comparatively low mark for "Advancement prospects" "The importance of jobs you are likely to do in the future in the RN" and "Your prospects in the RN". These young men perceived themselves to be categorised as drop outs and incompetents with no future in the service and this was by no means true. The solution is to make it clear that there is little or reflection on their competents, to ensure that being held ashore does not effect their advancement and if at all possible send them to shore billets where they can exercise their professional skills.

To summarise it is apparent that the results of a job analysis can be useful to the training design process from start to finish, but only if the designer is able to give clear direction to the analysts on the type of data he needs and the most useful ways of presenting the resulting information. If the job analysis is conducted without active involvement of the training designer only marginal use can be made of results because they are likely to be at an inappropriate level or in an inconvenient format. Job Analysis, particularly with the valuable CODAP routines, is part and parcel of training design and should so be used to provide a reasonable time scale cost - effective, objective courses.

The Army's Adoption of Performance Testing for MOS Proficiency

Captain Harry W. Porthouse
Training Management Institute
US Army Training & Doctrine Command

In the Enlisted Personnel Management System, a series of performance oriented tests called Skill Qualification Tests will replace the old MOS Test. The purpose of this paper is to describe that new testing system.

The most significant factor driving the Army toward Performance Testing for MOS proficiency has been the recognition by the Army leadership that retention and career progression should be based on demonstrated performance of critical job related skills. The MOS test, the only Army-wide proficiency testing system, has not highlighted skill deficiencies, shown the state of individual skill mastery in absolute terms, or given commanders an accurate picture of personnel readiness.

Historically, the Enlisted Evaluation Score (EES), composed of MOS test and efficiency report scores, was first used as a means of determining eligibility for award of proficiency pay in 1959. Since a fixed amount of money was available, the EES was devised as a normative system to allocate the dollar quota. Assessment of specialty knowledge for commissioned and warrant officers reverting to enlisted status was begun a year later, while verification of primary MOS (1961) and secondary MOS (1964) led to the use of the EES for promotion qualification in 1968.

1968 marked the year that the Profile Summary Report was introduced as the training proficiency indicator of the system. Other uses in the active Army were found over the years until a total of 15 programs are now dependent on the MOS test score. All of these programs are of the personnel management variety, the Profile Summary Report being the only feedback for training provided in the system.

Because of the personnel management focus of the overall system, various system components were undesirable from the training management point of view. First, Training and Doctrine Command schools, along with other agencies, were tasked to provide items for the MOS test, without influencing the objectives of the test.

The format for the MOS test was 125 multiple choice items. The response array for each item was constructed in the form of distractors in the traditional sense, meaning that the purpose of the item was to discriminate among examinees. If an item did not have sufficient discriminant validity, it was not reused. Test results were referenced to points on the normal curve, without regard to standards for task performance. In fact, very few MOS tests were able to deal with the ability to apply job knowledge in a performance setting. Since the purpose of the test was to discriminate, item writers tended to go farther and farther afield to discover the esoteric knowledge buried deep in technical manuals, and any possible correlation to critical job requirements was lost. This tended to reduce the value of the feedback for training in the Profile Summary Report.

With the advent of the new Enlisted Personnel Management System, the Army has committed itself to dual purpose tests. Beginning with the first Skill Qualification Tests in October 1976, proficiency testing must serve the needs of both the training and personnel management communities.

The guidance for the new SQT system as set forth by GEN DePuy, TRADOC Commanding General, on 1 August 1975, stated that the SQT must meet the criteria shown in this slide (Slide 1 on). As you can see, this will allow the evolution of a testing system solving many of the problems of the old MOS test. First, by basing the SQT on the critical tasks of the MOS, the soldier is not misdirected to less relevant aspects of his job. These critical tasks will be communicated to soldiers, supervisors, and commanders by means of the Soldier's Manual for the specific skill level of an MOS. The recording of the critical tasks, including job conditions, standards and measures of performance, will allow trainers at all levels to focus on the essentials of the MOS as they relate to success in combat, in shops, and in offices. (Slide 1 off)

The way in which skill levels of an MOS translate into an SQT progression pattern is shown in this slide (Slide 2 on). Attainment of a lower passing score tells a soldier that he has verified his present skill level. If the higher passing score is achieved, he is qualified for award of the next higher skill level. Award of the higher skill level is one prerequisite for promotion. (Slide 2 off)

At this point, let me relate some of the general characteristics of the SQT (Slide 3 on). As you can see, there are two basic models. The one on the left is for the lower levels (2 and 3), while the one on the right is for the higher levels (4 and 5). The difference between the two models lies in the hands-on component. It is felt that the nature of jobs tested by SQT 4 and 5 (pay grades E6 through E9) often limits performance testing. Reliable and valid observation of leadership and supervisory skills is difficult at best. The diagonal line you see in the model on the left signifies that, for the lower skill levels, the size of the hands-on component will increase over time. There are two reasons for this. First, in line with GEN DePuy's guidance, we did not want resource requirements for the portion of the new test administered in the unit to be perceived as excessive. Second, until we have gained experience in the standardized administration of performance tests

Army-wide, the TRADOC wants to hold down the size of this component until its results can be assessed. The size of the hands-on component for the higher skill levels will remain constant over time. You will hear more in the next paper to be presented about the research currently underway on methods to minimize potential problems in this area.

Another important characteristic of the SQT is the quarterly test administration concept. Currently, the Enlisted Evaluation Center specifies the month in which the MOS test will be taken. For the SQT, a test quarter will be specified. Each installation may select any month of that quarter as the month in which the SQT will be conducted for the MOS. The TRADOC had hoped that the SQT could be spread over the entire quarter at an installation, but problems with test compromise, collating results of the three components of the test, and scoring lag times preclude that approach for now.

Each installation within the United States and each division overseas will be responsible for conducting the SQT for all soldiers within a specified geographical area. Maximum use will be made of the existing Test Control Officer network to assure that soldiers are given the opportunity to take all components of the SQT. A standardization formula has been derived for those cases where soldiers are unable to take some portion of the test through no fault of their own. When the responsible installation delegates authority for the administration of the hands-on component, the installation or division is still responsible for standardization of test sites and administration, as well as for certifying that the test has been conducted in an honest manner.

As you can see from the slide, we will select more tasks from the lower skill level in each SQT. This is in recognition of the cumulative nature of most Army jobs. As a soldier moves from one skill level to another, the usual case is that only a few more tasks must be mastered, while he is still responsible for tasks at the lower skill level. In that respect, it is possible to infer some measure of success at a higher skill level from success at the lower skill level (Slide 3 off).

As stated earlier, and as shown by the previous slide, the SQT may have three components. I would like at this time to discuss each component briefly. (Slide 4 on) The task certification component is new to the Army's testing system, although it has been used for many years by some of the other services. This component, supplementing the hands-on tests, gives us a means of evaluating tasks and events that could not otherwise be tested in the SQT.

An example of a task which may be certified is "grade a road-bed." Equipment and other resource constraints make it impossible to test this task and others of high psychomotor content adequately in a formal performance test setting. However, we feel it is both possible and desirable to measure ability in areas such as these, using the performance measures and standards in the Soldier's Manual. Task certification also makes it possible to give the soldier credit for individual weapons qualification and physical fitness tests, and it may well be that new more relevant tests in these areas will result from including them in the task-based test. (Slide 4 off, Slide 5 on) By identifying tasks to be certified in the Soldier's Manual, and by giving supervisors 12 months in which to certify the tasks, major resource problems in the SQT are avoided. Because we recognize that certification may require more controls than are provided in the Soldier's Manual, no more than 10 tasks are to be certified in an SQT. We are currently examining whether certification is an appropriate alternative to testing per se in low density MOS where it is not cost effective to develop a full SQT. (Slide 5 off)

The written component of the SQT is also new. It is like the old MOS test only in that this component may present stimuli to an examinee in a paper and pencil format (or by audio-visual means) and calls for the examinee to produce machine scorable responses. As stated earlier, (Slide 6 on) all problems or questions must be performance relevant. Simulation is being encouraged by clustering the spaces available on the response sheet. For simulation, the 1 to 10 question spaces, each consisting of up to 10 response option spaces, can be converted to a matrix of up to 100 responses to a situation (10 x 10). This will be especially useful for tasks requiring review of elements of a product (such as a form) by the examinee.

Further, when purely knowledge items are used, they will be treated quite differently from the discrete items in the MOS test. The questions asked, or problems posed, to test knowledge of a task will be grouped and scored by task rather than item by item. In other words, if ten knowledges are requisite to performing a task, 8, 9 or even all 10 knowledge items may have to be answered correctly before credit will be awarded for the task (or scorable unit). In recognition of the inferential problem this poses, namely that the application of that knowledge in task performance is not tested, we see the written component of SQT 2 and 3 being decreased in time as the value of the hands on tests becomes apparent Army-wide and acceptance grows. (Slide 6 off)

Let us consider for a few moments the component to be made up of the hands-on tests. (Slide 7 on) Starting small, as you can see, SQT 2 and 3 will only test 4 to 16 tasks hands-on, while SQT 4 and 5 may contain anywhere from 0 to 10 hands-on tests. This is to insure that the component can be completed in four hours by an examinee. Test development agencies must also ascertain that the component can be administered installation-wide in a reasonable period. Two days are allowed for all except high density (e.g., combat arms) SQT, which cannot exceed five days at any post. The two principle resource constraints in hands-on testing are the ones you would expect: the availability of equipment and knowledgeable evaluators. Reliability and standardization are also considerations to be dealt with. The use of audio-visual media, both within the test and to present instructions to examinees and examiners, is being encouraged in order to come to grips with these problems. (Slide 7 off)

Another characteristic of the hands-on component which I would like to discuss (Slide 8 on) is what we are calling the Mandatory Task Concept. Certain tasks have been found to be crucial to the operational effectiveness of weapons or other critical systems. The state of soldier proficiency on these tasks will probably be proven to be less than desired, or even totally unacceptable. It is for this reason that certain tasks will be mandatorily included in the SQT every year. This establishes accountability for the soldier. His Soldier's Manual will give him the means to master the task, and the SQT will call him to account. Since we already know that performance and training for those tasks is inadequate, it would do little good and considerable harm to fail soldiers on the SQT for failing mandatory tasks initially. However, we feel that mandatory inclusion of these tasks in the SQT will provide sufficient training leverage to make such a stipulation possible at a later date. Naturally, such leverage would be lost if many tasks were declared mandatory, so we have introduced two control elements. First, no authority below TRADOC may designate a task as mandatory; second, no more than 5% of the SQT may be composed of mandatory tasks. (Slide 8 off)

Because scoring of the SQT will be covered in another paper today, I will limit my discourse on that topic to the basic scoring (Slide 9 on) rules. First, most tasks and knowledges that are tested or certified will be treated as one scorable unit. This was discussed earlier in the section on the written component. Extremely difficult, long, or complex tasks may be divided and treated as two or there scorable units. This in effect prevents underweighting these tasks--getting only one point for a time-consuming task. Additional points can also be granted for tasks and test events which have clearly definable, meaningful, well-known levels of proficiency (standards); for example, physical fitness, marksmanship, typing, and receiving or sending Morse Code. As I have inferred, each scorable unit is scored as go or no go. Finally, receiving a go on a task is dependent on the standard for the task or on the expert judgment of the test developers when standards are not explicit. In both cases, the possibility of measurement error must be considered when setting the test standard. The question of what standard to set must be addressed in the three validation phases of the SQT, as you will hear in the next presentation. (Slide 9 off)

I believe you can see by this time that the SQT is very different from the MOS test that it replaces. The system which supports the SQT will also be different. My presentation will conclude with its description.

In May of 1974, the Army Chief of Staff directed that the MOS testing function be transferred from the Military Personnel Center to TRADOC. It has now been agreed by both parties that this transfer will occur in accordance with the EPMS Implementation Schedule. This means that the Enlisted Evaluation Center will maintain the MOS tests, with the assistance of the Item Writing Agencies, until they are phased out incrementally through 1979. TRADOC will guide the development of the SQT as they are phased in during that same period.

To do this, (Slide 10 on) the Enlisted Test Activity has been established at Fort Eustis, VA. The functions of the ETA are to: develop and recommend SQT policy and procedures, design and implement the SQT system, establish SQT requirements for Test Development Agencies, supervise the validation of SQT, act as approval authority for SQT, produce and distribute the SQT, process and analyze SQT data, and formulate and distribute SQT reports.

One of the aims of the ETA is to reduce the processing time for test results. After the test is taken, the data will be transmitted to the ETA, processed into test results there, and the results reported to the individual, various levels of command, test developers and personnel managers. Whether this transmission will be done by AUTODIN, as shown here, or by other means is currently under study. The reason for attempting to reduce processing times is to provide quicker feedback to the individual and his commander. Further, since the SQT is task-based, feedback is highly specific, allowing prescriptive action which aids professional development and unit readiness. (Slide 10 off)

This paper has attempted to portray the fact that, although SQT represents a major change in Army proficiency testing, we are continuing to evolve the Enlisted Evaluation System begun in 1959. New methods of measuring job skills for training and personnel management purposes are constantly being sought, and will obviously cause further change to the system in coming years. As you will see in the other two papers in this program, procedures are being considered, detailed implementation planning carried out, and issues discussed. But the day when Army-wide performance testing becomes common practice is fast approaching.

CG's GUIDANCE FOR SQT

GENERAL

STARK SIMPLICITY

EVOLUTION, NOT REVOLUTION

SPECIFIC

CRITICAL TASK BASED

PERFORMANCE RELEVANT

MODEST, PRACTICAL HANDS-ON COMPONENT INITIALLY

IMPROVE WRITTEN TEST

INCLUDE PERFORMANCE CERTIFICATION

GOAL

DEVELOP SQT THAT WILL --

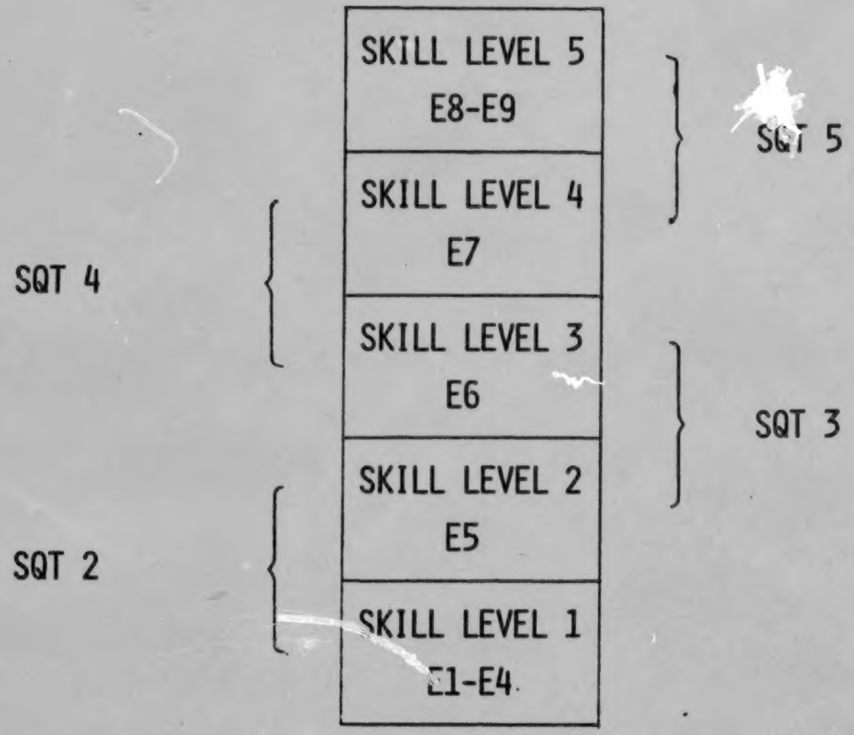
---MEASURE THE PROFESSIONAL COMPETENCE OF THE SOLDIER

---SERVE THE NEEDS OF PERSONNEL MANAGEMENT

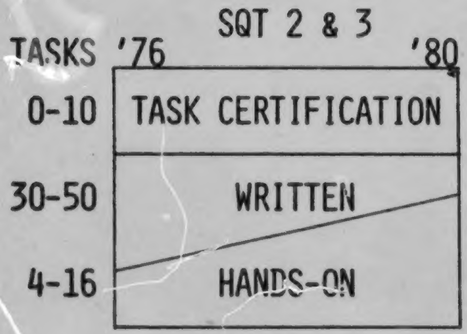
---BE PERCEIVED BY THE SOLDIER AS FAIR AND RELEVANT

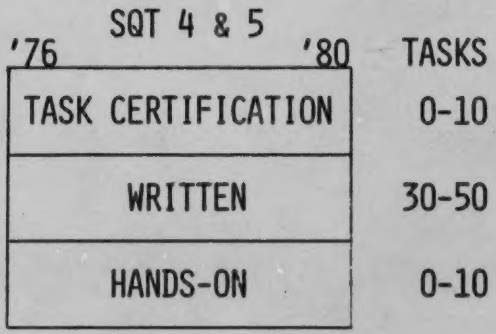
---NOT CONSUME AN UNACCEPTABLE AMOUNT OF UNIT RESOURCES IN ITS CONDUCT

SQT PROGRESSION


SKILL LEVEL 5
E8-E9
SKILL LEVEL 4
E7
SKILL LEVEL 3
E6
SKILL LEVEL 2
E5
SKILL LEVEL 1
E1-E4
SQT 5
SQT 4
SQT 3
SQT 2

GENERAL CHARACTERISTICS OF SQT

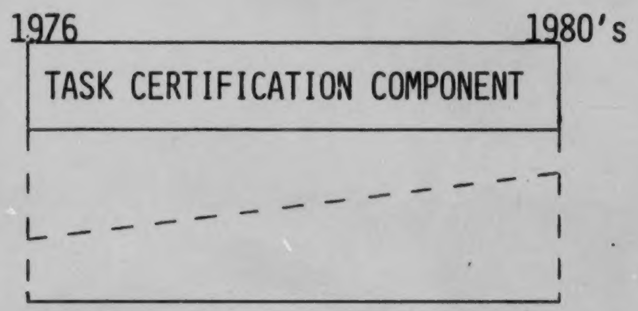

SQT 2 & 3
TASKS
'76
'80
0-10
TASK CERTIFICATION
30-50
WRITTEN
4-16
HANDS-ON


SQT 4 & 5
'76
'80
TASKS
TASK CERTIFICATION
0-10
WRITTEN
30-50
HANDS-ON
0-10

- SQT MAY HAVE THREE COMPONENTS.
- SAMPLES FROM CRITICAL TASKS.
- MAXIMUM OF 8 HOURS REQUIRED TO TAKE TEST.
- RESPONSIBILITY FOR CONDUCT FIXED AT INSTALLATION (CONUS) AND DIVISION (OCONUS) LEVEL.
- EACH INSTALLATION/DIVISION RESPONSIBLE FOR ALL MOS INCUMBENTS WITHIN A SPECIFIED GEOGRAPHIC AREA.
- TEST WILL BE CONDUCTED WITHIN A SPECIFIED CALENDAR QUARTER.
- WRITTEN AND HANDS-ON COMPONENTS WILL BE CONDUCTED WITHIN THE SAME CALENDAR MONTH OF THE SPECIFIED CALENDAR QUARTER.
- 2/3 TO 3/4 OF TASKS IN A GIVEN SQT WILL MEASURE SOLDIER'S COMPETENCE IN HIS PRESENT SKILL LEVEL.
- 1/4 TO 1/3 OF TASKS IN A GIVEN SQT WILL MEASURE SOLDIER'S COMPETENCE IN THE NEXT HIGHER SKILL LEVEL.
- FOR MOST MOS THE HANDS-ON COMPONENT OF SQT 2 & 3 WILL EXPAND OVER TIME ACCOMPANIED BY A DECREASE IN THE WRITTEN COMPONENT.

CHARACTERISTICS OF TASK CERTIFICATION COMPONENT

SQT 2 & 3



- NATURE OF TASKS CERTIFIED:
  - IMPRACTICAL TO TEST HANDS-ON IN SQT
  - ADEQUATE SIMULATION NOT AVAILABLE
  - HAVE HIGH SKILL CONTENT
- ADMINISTERED AS SUPPLEMENT TO HANDS-ON COMPONENT.
- SCORING OPTIONS:

*NO GO (PERFORMANCE INADEQUATE)	0
*GO (PERFORMANCE ADEQUATE)	1
GO (PERFORMANCE EXCEEDS MINIMUM STANDARDS)	2
GO (PERFORMANCE OUTSTANDING/WARRANTS SPECIAL RECOGNITION)	3
*NOT OBSERVED	✓

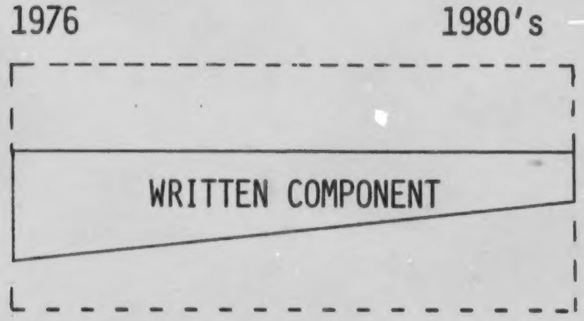
*NOTE: THESE OPTIONS APPLY TO ALL CERTIFIED TASKS; THE REMAINING OPTIONS APPLY ONLY WHERE WARRANTED, E.G., INDIVIDUAL WEAPONS QUALIFICATIONS AND PHYSICAL FITNESS TEST.

CHARACTERISTICS OF TASK CERTIFICATION COMPONENT (CONT)

- COMMANDER/SUPERVISORS WILL BE ENCOURAGED TO OBSERVE TASKS TO BE CERTIFIED USING THE SOLDIER'S MANUAL TO IDENTIFY PERFORMANCE MEASURES AND STANDARDS.
- TASKS TO BE CERTIFIED MUST BE OBSERVED WITHIN 12 MONTHS PRECEEDING THE TEST QUARTER.
- NO MORE THAN 10 TASKS FOR CERTIFICATION WILL BE INCLUDED IN AN SQT UNLESS THE ENTIRE SQT CONSISTS OF PERFORMANCE CERTIFICATIONS (LOW DENSITY MOS).

CHARACTERISTICS OF WRITTEN COMPONENT

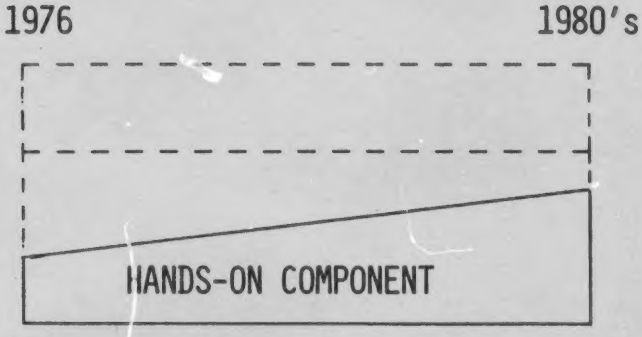
SQT 2 & 3



- TESTS 30-50 TASKS (SCORABLE UNITS).
- 1-10 PROBLEMS/QUESTIONS ALLOWED PER TASK.
- 2-10 RESPONSES ALLOWABLE PER PROBLEM/QUESTION.
- ALL PROBLEMS/QUESTIONS MUST BE PERFORMANCE RELEVANT.
- PROBLEMS WILL SIMULATE PERFORMANCE WHERE FEASIBLE.
- 90% MUST COMPLETE IN 4 HOURS.
- SOLDIER PRODUCES MACHINE-SCORABLE RESPONSES.
- EXPERT EVALUATOR NOT REQUIRED.
- ADMINISTERED TO GROUPS BY TCO.
- COMPONENT IS SELF-CONTAINED EXCEPT WHEN ELECTRONIC MEDIA USED.

CHARACTERISTICS OF HANDS-ON COMPONENT

SQT 2 & 3



- TESTS 4-16 TASKS (SCORABLE UNITS).
- 1-20 PERFORMANCE MEASURES ALLOWABLE PER TASK.
- EACH TASK HAS TIME STANDARD.
- WILL REQUIRE NO MORE THAN 4 HOURS OF EXAMINEE'S TIME.
- EXCEPT FOR HIGH DENSITY MOS (11B, ETC.), COMPONENT MUST BE ADMINISTRABLE TO ENTIRE MOS POPULATION IN 2 DAYS AT A TYPICAL INSTALLATION (DIVISION).
- COMPONENT MUST BE ADMINISTRABLE TO HIGH DENSITY MOS IN 5 DAYS AT A TYPICAL INSTALLATION.
- EXPERT EVALUATORS ARE REQUIRED.
- PROCESS AND/OR PRODUCT WILL BE SCORED.
- EVALUATOR PRODUCES MACHINE-SCORABLE RESPONSES.
- USE OF SIMULATION AND AUDIO-VISUAL MEDIA IS ENCOURAGED.
- AIDS AVAILABLE ON THE JOB WILL BE AVAILABLE FOR THE TEST.

MANDATORY TASK CONCEPT

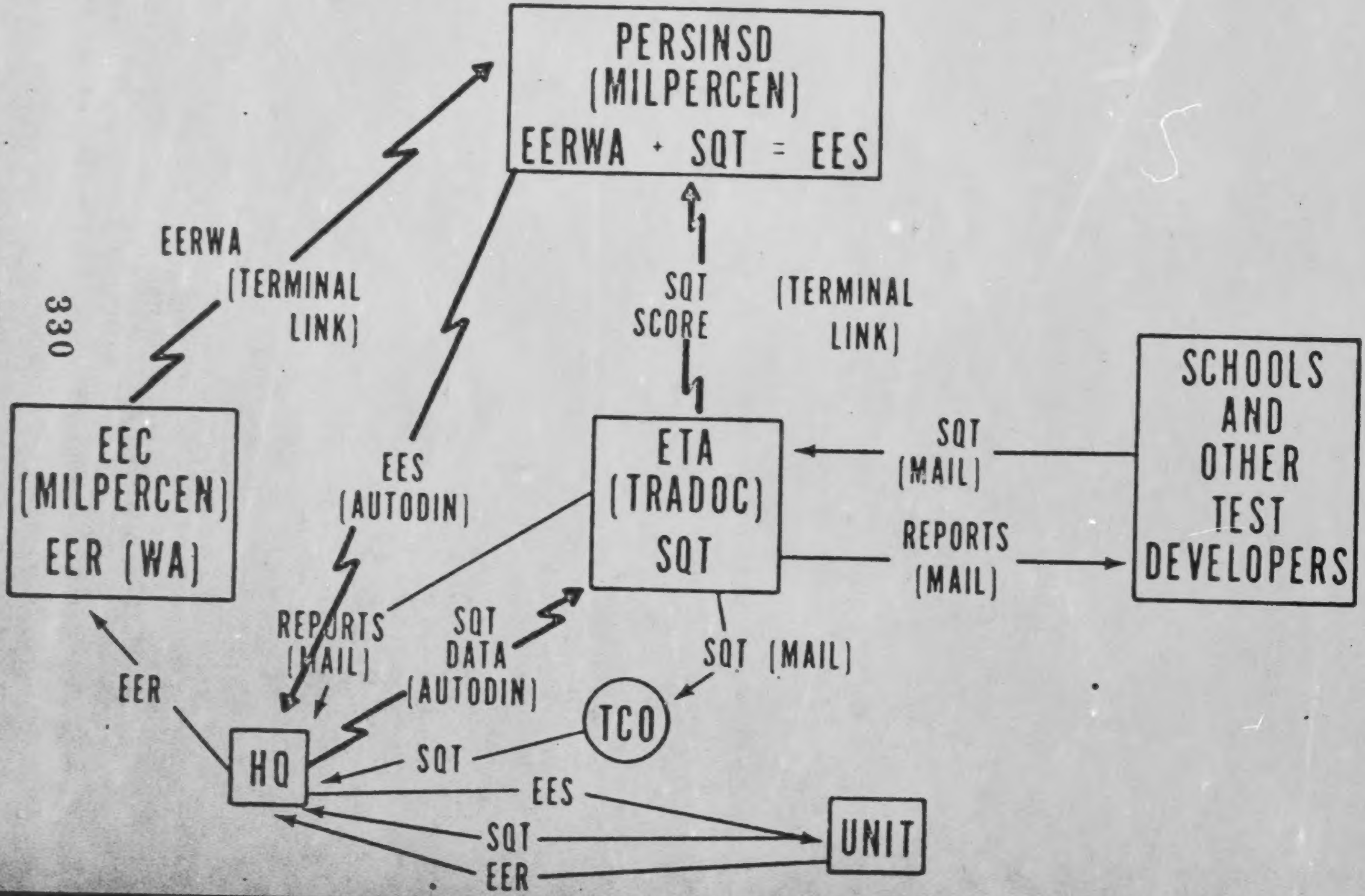
CHARACTERISTICS OF MANDATORY TASK

- BEARS DIRECTLY AND SIGNIFICANTLY ON THE OPERATIONAL EFFECTIVENESS OF WEAPONS OR OTHER CRITICAL SYSTEMS.
- IDENTIFIED AS MANDATORY IN SOLDIER'S MANUAL.
- TESTED IN ALL SQT FOR MOS TO WHICH IT APPLIES AS LONG AS IT IS MANDATORY.
- NO AUTHORITY BELOW CG TRADOC MAY DESIGNATE A TASK MANDATORY.
- THE NUMBER OF TASKS DESIGNATED MANDATORY SHOULD NOT EXCEED 5% OF THE TOTAL CRITICAL TASKS UPON WHICH AN SQT IS BASED.
- TESTED IN HANDS-ON COMPONENT IF SUCH EVALUATION IS PRACTICAL.
- A GIVEN MANDATORY TASK MAY APPLY TO ONE OR MORE MOS.
- A PROPONENT MAY BE DESIGNATED FOR A MANDATORY TASK APPLICABLE TO SEVERAL MOS TO DEVELOP SOLDIER'S MANUAL TASK DESCRIPTION, TRAINING MATERIALS AND SQT EVENT (SCORABLE UNIT).

SQT SCORING RULES

- MOST TASKS WILL BE TESTED/CERTIFIED AND SCORED AS ONE SCORABLE UNIT.
- EXTREMELY COMPLEX AND DIFFICULT TASKS MAY BE BROKEN INTO SUB-TASKS AND SCORED AS 2-3 SCORABLE UNITS.
- TASKS HAVING DEFINABLE, MEANINGFUL AND SCORABLE LEVELS OF PROFICIENCY (STANDARDS) MAY BE SCORED AS 2-3 SCORABLE UNITS TO ALLOW EXTRA CREDIT FOR THE MORE SKILLED SOLDIER (E.G., AAQ, PHYSICAL FITNESS TEST, TYPING, SENDING/RECEIVING MORSE CODE, ETC.).
- EACH SCORABLE UNIT IS SCORED GO OR NO GO.
- EACH PROBLEM/QUESTION WITHIN A SCORABLE UNIT IS SCORED GO OR NO GO.
- SCORING CRITERIA FOR SCORABLE UNITS AND PROBLEMS/QUESTIONS WITHIN SCORABLE UNITS WILL BE BASED ON EXPERT JUDGMENT MODIFIED FOR MEASUREMENT ERROR.

PROPOSED ENLISTED EVALUATION SYSTEM



PERSINSD
(MILPERCEN)
EERWA + SQT = EES

EERWA
(TERMINAL
LINK)

SQT
SCORE

(TERMINAL
LINK)

EEC
(MILPERCEN)
EER (WA)

EES
(AUTODIN)

ETA
(TRADOC)
SQT

SQT
(MAIL)

REPORTS
(MAIL)

SCHOOLS
AND
OTHER
TEST
DEVELOPERS

REPORTS
(MAIL)

SQT
DATA
(AUTODIN)

SQT (MAIL)

EER

TCO

HQ

SQT

EES

SQT

EER

UNIT

CURRENT AND PROJECTED RESEARCH IN SKILL QUALIFICATION TESTING

Douglas L. Young, Ph.D.
U.S. Army Research Institute

Today is clearly an appropriate time to describe some of the research on skill qualification testing. This research encompasses a number of interrelated efforts: field study of performance testing, SQT performance test development and workshops, SQT design and validation, feasibility and acceptability checklists, and SQT Two in six Combat Arms MOS.

FIELD STUDY OF PERFORMANCE TESTING

Under contract to the Army Research Institute, HumRRO conducted a study to determine the feasibility of administering and scoring performance tests with untrained examiners. Four tests were prepared by the U.S. Army Infantry School. Five groups of five NCOs, with one serving as the test administrator, observed fifteen soldiers and scored them on each test. More methodological detail will be available in the project final report which is presently under review. Inter-rater reliability estimates indicate tests could not be scored reliably because thirty-five percent of the performance measures within tests were not scored reliably. Even this weak degree of relationship may have been enhanced by a very low GO rate. On the average, eighty percent of the soldiers failed at least one of the performance measures composing each test. In three of the four tests, a soldier must pass all performance measures in order to receive a GO on the test. Moreover, considerable test site variance was noted; i.e., testing conditions were unstandardized. In addition some critical performance measures were omitted or not scored. In sum, performance tests are difficult to administer and score in a standardized manner; therefore, attenuation of validity and reliability is a serious problem when using performance tests in the field.

The views expressed in this paper are those of the author and do not necessarily reflect the views of the U.S. Army or the Department of Defense.

Requests for reprints should be sent to Douglas L. Young, U.S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army, 1300 Wilson Boulevard, Arlington, Virginia 22209.

SQT PERFORMANCE TEST DEVELOPMENT AND WORKSHOPS

Some efforts have been made to reduce the problem of validity and reliability restriction when using performance tests in the field. Among these efforts is the manual on performance test development for skill qualification testing by Vineberg and Taylor (1975). Recently TRADOC has placed less emphasis on performance testing at the lower skill levels in favor of performance and knowledge tests that are paper and pencil administrable. A revised version of the manual will address the following topics: task sampling, development of hands-on and written tests, and scoring.

The manual and a SQT design and validation plan, which will be presented later, are being used in a series of four workshops. These workshops provide guidance for the development of SQTs to personnel responsible for test construction in the various Career Management Fields. Concepts of both written and hands-on test construction are presented and discussed. Practical exercises are conducted on content examples from various MOS; through critiques of practical exercises, assistance is provided in solving problems of test development. Most of the workshop is devoted to small group work on their own particular problems and subject matter content; however, the small groups do come together near the end to report and discuss what happened in each of the small groups. In sum, technical expertise has been brought to bear on the problems of reliability and validity by constructing a manual and conducting workshops which attend to the problems.

SQT DESIGN AND VALIDATION

A plan for designing and validating SQTs was presented at the EPMS Implementation Conference, 13 August 1975. This aid was revised and expanded for use at the workshops. The 4 September revision is attached; major steps of SQT design and validation follow.

1. Identify the critical and mandatory tasks of each skill level. Proponent agencies identify critical tasks while TRADOC will designate as mandatory tasks which must be included in an SQT. Additional tasks may be nominated for certification by commander or supervisor; but this group of tasks or component will not be discussed.

2. Select task content of the SQT. More than forty-five common or key duty position tasks should be selected. All critical tasks are included in the Soldier's Manual. In all probability there will be too many mandatory and critical tasks to test in the SQT; therefore, sampling may be necessary.

3. Design an ideal or full performance test for each task. These task tests have two purposes: they can be used in the hands-on component, if resource demands are not excessive; or if demands are excessive, performance tests can be used to validate items in the written component.

4. Construct task tests for hands-on component, including the mandatory tasks whenever possible. Up to four hours may be used in hands-on testing. Equity and standardization are required of all tests. Skilled tasks, which consume a large amount of training and practice time, are prime candidates for inclusion as hands-on tests.

5. Determine resource and practical constraints in field units. Considerations include: MOS concentrations, testing site locations, numbers of qualified evaluators available at sites, and amounts of equipment and type of terrain at sites. If a task test requirements exceed testing units' capabilities, the test cannot be included in the hands-on component.

6. Develop a scorable unit for each task. In the hands-on component each performance test of a task is scored zero or one. In the written component each test of a task is also scored zero or one; however, more than one item per task may be used to estimate mastery of the task.

7. Validate scorable units (Phase I Validation) on local task masters and nonmasters. Inter-rater reliability and discriminant validity of hands-on tests will be assessed. Written items will be validated against performance tests of the same task whenever possible. Remaining written items will be validated against all other items from the same functional area. Acceptability or face validity will be evaluated informally.

8. Validate both components (Phase II Validation) with local representative personnel. Inter-rater reliability of tests in the hands-on component is determined, and item-task agreement is calculated for items in the written component. Acceptability of scorable units in both components is assessed by using checklists. After unacceptable and/or invalid items are rejected, the written component will be rescored. Then administrative feasibility of both components will be determined.

9. Transmit information to the TRADOC Training Management Institute (TMI). The sampling plan which lists tasks selected for hands-on and written components as well as validation of their scorable units should be sent with the results of Phase II Validation indicating: representativeness of examinees, difficulty indices, validity indices, standardized administration, acceptance, and feasibility. Also a copy of the complete components should be transmitted for TMI review.

10. Pilot test hands-on component (Phase III Validation) in FORSCOM, USAREUR, and RESERVE or GUARD Units. This pilot test assesses acceptability and feasibility in field units. Resulting data will be transmitted to TMI.

FEASIBILITY AND ACCEPTABILITY CHECKLISTS

Checklists are being devised by the Army Research Institute for field trials of both written and hands-on task tests. One type of checklist will be appropriate for examinees, another for test administrators, scorers, as well as supervisors, commanders, or training managers. Checklists attend to various aspects of tests that are potentially troublesome: (a) characteristics of tasks [e.g., importance, performance frequency, and relevance]; (b) characteristics of test administration [e.g., instructional clarity, site feasibility, safety, conditions, and standards]; and (c) characteristics of scoring [e.g., clarity, appropriateness, and demands]. A summary form will be provided to use with the checklists which lists the characteristics, question number, and favorable response to each question. Spaces will be marked for the number and proportion of responses that can indicate shortcomings or deficiencies and the need for revision.

SQT TWO IN SIX COMBAT ARMS MOS

Group I (Career Management Fields 11, 16, 63, 74, and 95) commences field validation of SQT Two in December 1975 with completion scheduled for March 1976. The Army Research Institute has selected six MOS (11B, 11C, 11D, 11E, 16D, and 16E) for close experimental attention. Proponent agencies for these MOS will go through design and validation steps similar to those presented earlier. For example, the Armor School has identified 77 and 26 critical tasks in skill levels one and two respectively. Table 1 presents the number of common Combat and Armor tasks as well as the driver, loader, and gunner key duty position tasks. The numbers of tasks selected for the written and hands-on SQT components are indicated for each category. These categories are mutually exclusive, thus the skill level one categories cover the domain of entry level Armor Crewman tasks except for duty positions which are not in the core track.

Full performance tests exist for all 103 tasks and field constraints are being determined. It is expected that the components encompassing thirty written and eighteen hands-on scorable units respectively can each be tested within four hours. The three phase validation of scorable units and components remains to be done.

TABLE 1

NUMBER OF ARMOR COMMON AND KEY DUTY POSITION TASKS SELECTED FOR WRITTEN AND HANDS-ON COMPONENTS

Categories	SQT Component		Unselected	Sum
	Written	Hands-on		
SKILL LEVEL ONE TASKS				
Common Combat	2	1	21	24
Common Armor	5	2	3	10
Driver Duty Position	4	5	14	23
Loader Duty Position	7	6	7	20
SKILL LEVEL TWO TASKS				
Gunner Duty Position	12	7[a]	7	26
TOTAL	30	21	52	103

[a] Three of these tasks involve live fire of the main gun; therefore, they will be Commander certified after Annual Gunnery.

One of the written scorable units developed from task analysis is "respond to visual signals". This task is performed under four sets of conditions: ground guide-day, ground guide-night, platoon leader-day, and platoon leader-night. Fortunately, the standards and expected behaviors do not change as a function of the conditions. Items will be written which evaluate the soldier's ability to recognize various signals when presented pictorially. In addition, the soldier's ability to give ground guide signals and respond to ground guide signals while driving during daylight can be assessed in the hands-on component. One soldier will give signals to another in a tank that he should: start tank engine, move forward, move backwards, stop tank, and stop tank engine. Then the soldiers can reverse roles.

In closing, research efforts in several related areas were presented. A field study of performance testing indicated problems of reliability and validity. Solutions to the problems were explored in a manual for SQT development and were discussed at workshops. Some details on SQT design and validation were presented, followed by feasibility and acceptability checklists. Finally some examples were drawn from the Armor crewman SQT Two activities which are underway.

SKILL QUALIFICATION TEST DESIGN AND VALIDATION BY PROPONENT AGENCIES
DOUGLAS L. YOUNG, ARMY RESEARCH INSTITUTE, 4 SEPTEMBER 1975

1. Identify the critical and mandatory tasks of each skill level. See TRADOC letter of 3 March 1975, Subject: SQT Two for Combat Arms, paragraph 2 for definitions of critical tasks. Mandatory tasks must be included in an SQT; they will be initially selected by the proponent school and then evaluated for appropriateness as manadatory by TRADOC Headquarters. Up to ten additional tasks may be selected for certification by commander or supervisor. The certification component is not discussed below.

2. Select content of the SQT, which consists of mandatory and critical tasks from the lower skill levels plus those from the next higher skill level. More than forty, preferably fifty or more, tasks common to all duty positions and critical tasks of the key duty positions should be selected. If the number of tasks is too large, arrange the tasks by functional areas and obtain a representative sample of more than forty tasks (See SQT Manual, Ch. 2). All critical tasks, regardless of the duty position in which they are performed, will be included in the Soldier's Manual and training development process. Critical tasks are eligible for inclusion in the SQT; however, prime consideration should be given to common and key duty position tasks.

3. Design an ideal or full performance test (See SQT Manual, Ch. 3) for each task. All performance tests must have high correspondence with job tasks to insure accurate measurement of skill level proficiency. Full performance tests have two purposes: they can be used in the hands-on component if the resource demands are not excessive (such as, soldier's time, evaluator's time, equipment, terrain, etc.); or if resource demands are excessive, full performance tests can be used to validate items of the written component.

4. Construct tests of tasks to be included in the hands-on component. If at all possible, mandatory tasks should be tested in the hands-on component; that is, by requiring the soldier to demonstrate his proficiency just as though he were performing on the job. A manual, developing Criterion-Referenced Tests by Swezey and Pearlstein, provides excellent guidance for constructing hands-on tests.

a. Up to four hours may be used for "hands-on" testing. Probably six to eighteen tasks can be tested hands-on within four hours. Resource demands of the SQT hands-on component must not exceed field unit capabilities. Hands-on performance tests must satisfy the requirements of equity and standardization. Each of these tasks must also be appropriate for administration in the National Guard or Army Reserves.

b. Skilled tasks, those which require a great deal of training and practice time, are prime candidates for inclusion as performance tests. Skilled tasks tested in the hands-on component must be the essence of the speciality, undoubtedly critical, common and key duty position tasks of the skill level.

5.. Determine or accurately estimate resource and practical constraints operating in field units. Questions such as the following must be answered to insure that an SQT can be administered in a standarized manner to provide equitable test scores.

a. Where are the soldiers of an MOS concentrated?

b. Where will the SQT testing sites be located?

c. What numbers of soldiers will be tested?

d. How many MOS-qualified evaluators (examiners and administrators) will be available at the SQT testing sites?

e. What amounts of equipment and type of terrain are available at all SQT sites of the MOS?

Final SQT requirements must not exceed the testing units' capabilities. Both requirements and capabilities must be examined carefully. If the requirements exceed capabilities, then the task cannot be included in the hands-on component.

6. Develop a scorable unit for each task. For the hands-on component of the SQT each performance test of a task provides a unit scored zero or one. If products are scored, the tests may be readily administered to groups of soldiers. If processes are scored, the tests must be administered to individuals one at a time. For the written component of the SQT, each test of a task also provides a unit scored zero or one; more than one item per task may be included to provide a pass/fail score for each task. The entire written component will be administered by the Test Control Officer.

7. Validate tests of tasks. Phase I validation is conducted locally by proponent agency.

a. Select a minimum of five masters and five nonmasters for each task. Ideally the masters and nonmasters should be comparable in all respects except for ability to perform the specific task; they should be comparable in terms of age, Army experience, and educational level. Also, ideally different groups of masters and nonmasters should be selected for each task. If the master and nonmaster groups are very different in terms of experience, age, or education, then the validation results may be of questionable value to the test developer. In actual practice, the same groups could be used to validate several tasks that are quite similar, such as masters from the same functional area.

b. Tasks in hands-on component.

(1) Two or more experts develop procedures for administering and scoring the scorable unit. The scorable unit is refined until the experts agree perfectly in scoring. They may administer the test to each other or to another individual. The scoring of the expert is the standard in subsequent steps of validation.

(2) Administer the hands-on task to groups of masters and nonmasters by one of the experts. Performance of each examinee is scored by an expert, who serves as the standard, and by at least three independent raters. Each step or observable element of the task is to be scored by the expert and all raters.

(3) Prepare the following type of table for each step of a task. Std = expert evaluator, A = agreements, and D = disagreements with standard. The table shows the minimum acceptable standards for evaluating each step of a hands-on task.

Master		Raters				
Examinees	Std	1	2	3	A	D
1	1	1	1	1	3	0
2	1	1	1	1	3	0
3	1	0	1	1	2	1
4	1	1	0	0	1	2
5	0	0	0	0	3	0
					12	3

Nonmaster		Raters				
Examinees	Std	1	2	3	A	D
1	1	1	1	0	2	1
2	0	0	1	0	2	1
3	0	0	0	0	3	0
4	0	0	1	0	2	1
5	0	0	0	0	3	0
					12	3

Agreement between the standard and raters must be eighty percent or higher for the master group and for the nonmaster group; that is, the evaluators must be able to score accurately both masters and nonmasters.

An additional requirement is that eighty percent or more of the masters must be GO on the total task, while twenty percent or less of the nonmasters are GO on the total task (80% or more of the nonmasters must be NO-GO on the total task).

(4) If the above minimum requirements are not met, the steps must be revised, and tried out until the total test of a task is satisfactory.

c. Tasks in written component, validated against a performance-oriented version of the task.

(1) Two or more experts develop procedures for administering and scoring a hands-on version of the task. The procedures are refined until the experts agree perfectly in scoring. They may administer the test to each other or to another individual. The scoring of the experts is the standard in subsequent steps of the validation. (This procedure is identical to that of 7b(1) above.)

(2) Develop a set of written items for each task. More items should be developed than will be needed in the final SQT.

(3) Administer both the hands-on and written versions to groups of at least five masters and five nonmasters.

(4) The minimum acceptable standard for the hands-on version is that eighty percent or higher of the master group pass the test and that twenty percent or less of the nonmaster can pass the test (80% or more of the nonmasters must fail the task tests). No evaluators, other than the expert, are required to observe administration of the hands-on version.

(5) Obtain the extent of agreement between the hands-on version and each of the written items. The maximal (100%) and minimal (60%) degrees of acceptable agreement are shown in the following tables.

MAXIMAL

WRITTEN ITEM	HANDS-ON VERSION GO	NO-GO	
Pass	5	0	5
Fail	0	5	5
Total	5	5	10

	GO	NO-GO	
Pass	6	0	6
Fail	0	4	4
Total	6	4	10

	GO	NO-GO	
Pass	4	0	4
Fail	0	6	6
Total	4	6	10

MINIMAL

WRITTEN ITEM	HANDS-ON VERSION GO	NO-GO	
Pass	3	2	5
Fail	2	3	5
Total	5	5	10

	GO	NO-GO	
Pass	4	2	6
Fail	2	2	4
Total	6	4	10

	GO	NO-GO	
Pass	2	2	4
Fail	2	4	6
Total	4	6	10

Sixty percent or more of the scores must be in agreement; that is, pass both the hands-on version and the written item or fail both. A minimum of sixty percent agreement must be obtained for each written item. In other words, 6 out of 10 men would be in the upper left and lower right cells of the 2 by 2 table.

d. Tasks in written component, NOT validated against a performance-oriented version of the task.

(1) Develop a set of written items for each task in a functional area which has not been previously validated. Again more items should be developed than will be needed in the final SQT.

(2) Administer the written items from a functional area to groups of fifteen expected masters and fifteen expected nonmasters of the functional area.

(3) Determine the total number of items the man passed. Order all scores from highest to lowest. Select the ten highest and the ten lowest scoring soldiers and consider them to be masters and nonmasters of the functional area.

(4) Obtain the numbers passing and failing each written item from both the highest scoring and lowest scoring groups. If four to sixteen of the twenty soldiers pass any item, at least sixty percent agreement must be obtained. Percent agreement is calculated from 2 by 2 tables as it was in 7c(5). If more than sixteen or less than four soldiers pass any item, at least fifty-five percent agreement must be obtained.

8. Validate components. Phase II validation is conducted locally by proponent agency.

a. Administer hands-on and written components to about fifty examinees who are representative of the Army-wide population that will take the SQT.

b. Obtain interrater reliability (agreement) for hands-on component tasks. A minimum of four test evaluators must be involved in administering and scoring the hands-on component. One administrator must give the task to at least ten examinees, and the performance must be scored independently by at least three raters, unless circumstances prevent simultaneous rating by that many individuals. Prepare the following type table for each step.

	RATER					
EXAMINEE	ADM	1	2	3	Agree	Disagree
1	1	1	1	0	2	1
2	1	1	1	1	3	0
3	0	0	0	1	2	1
4	1	0	1	1	2	1
5	0	0	1	0	2	1
6	0	0	0	0	3	0
7	1	1	1	1	3	0
8	1	0	1	1	2	1
9	1	1	1	0	2	1
10	0	0	0	0	3	0
			TOTAL		24	6

% agreement =

$$\frac{24}{30} \times 100 = 80$$

The minimum requirement is eighty percent agreement of the raters with the administrator for each step. If this requirement is not met, revise the step and/or scorable unit.

c. Obtain the extent of agreement between pairs of items in the scorable units of the written components if there are four or fewer items in the unit. If there are five or more items in the scorable unit, obtain the extent of agreement between each item and pass-fail on the scorable unit. The comparisons for scorable units with two, three, four, five, and six items are as follows:

Two Item Scorable Unit (SU): Items 1&2;
Three Item Scorable Unit (SU): Items 1&2, 1&3, and 2&3;
Four Item Scorable Unit (SU): Items 1&2, 1&3, 1&4, 2&3, 2&4, and 3&4;
Five Item Scorable Unit (SU): Item 1&SU, 2&SU, 3&SU, 4&SU, and 5&SU;
Six Item Scorable Unit (SU): Items 1&SU, 2&SU, 3&SU, 4&SU, 5&SU, and 6&SU.

(1) In the case of scorable units which have less than five items, at least sixty percent agreement is required. Percent agreement is calculated from 2 by 2 tables as it was in 7c(5). For example the agreement is sixty percent in the following table:

	ITEM 2		
ITEM 1	Pass	Fail	
Pass	15	9	24
Fail	11	15	26
TOTAL	26	24	50

When agreement is less than sixty percent for pairs with a particular item, that item has some sort of shortcoming and must be examined for revision.

(2) In the case of scorable units which have five or more items, at least sixty percent agreement is required when item difficulty is between .20 and .80. Following are examples of 2 by 2 tables where the item has a difficulty of .20 and .80, respectively:

	SCORABLE UNIT		
ITEM	Pass	Fail	
Pass	5	5	10
Fail	15	25	40
TOTAL	20	30	50

	SCORABLE UNIT		
ITEM	Pass	Fail	
Pass	25	15	40
Fail	5	5	10
TOTAL	30	20	50

When item difficulty is less than .20 or more than .80, at least fifty-five percent agreement is required. Percent agreement is calculated from 2 by 2 tables (item by scorable unit) as it was in 7c(5). It is possible to adjust scorable unit difficulty by altering the standard and thus increasing present agreement.

Following are examples of 2 by 2 tables where the item has a difficulty of .10 and .90 respectively:

	SCORABLE UNIT		
ITEM	Pass	Fail	
Pass	3	2	5
Fail	20	25	45
TOTAL	23	27	50

	SCORABLE UNIT		
ITEM	Pass	Fail	
Pass	25	20	45
Fail	2	3	5
TOTAL	27	23	50

d. Determine acceptability of written and hands-on components as measures of MOS relevent tasks. Model questionnaires will be provided. Include examinees, supervisors, and commanders in survey.

e. Reset scorable unit standards of written component after rejecting unacceptable and invalid items, then rescore the component based on the revised version. Component will be evaluated by calculating percent agreement between scorable units and revised component total score. Order all examinees on their total score, then divide them into two equal sized groups. Obtain the numbers passing and failing each scorable unit from both the highest and lowest scoring groups. If the scorable unit difficulty is between .20 and .80, at least sixty percent agreement is required. When difficulty is less than .20 or more than .80, at least fifty-five percent agreement is required. See 8c(2) for examples. If these requirements are not met, scorable units should be carefully examined to determine if they have any deficiencies or shortcomings. Scorable units that do not meet the requirements may be retained if their inclusion can be justified on the basis of content validity, criterion-related or part-whole validity, and/or acceptability (face validity).

f. Determine percent of examinees that complete the revised written and hands-on components within time limits.

g. Determine feasibility of administering hands-on component in a field environment.

9. Provide data to Training Management Institute, Fort Eustis.

a. Prior to validation send Sampling Plan, which contains the following information.

(1) Rationale and tasks selected for SQT from each skill level.

(2) Tasks selected for hands-on component.

(3) Tasks selected for written component.

(a) Task tests validated against performance-oriented version of the task.

(b) Task tests not validated against performance-oriented version of the task.

b. Phase 2 Validation.

(1) Description of sample to determine representativeness of examinees.

(2) Difficulty of each item or step and scorable unit.

(3) Validity of each scorable unit.

(4) Summary of analysis and certification that minimal acceptable standards of test administration have been met.

(5) Evidence that written and hands-on components are acceptable to examinees, evaluators, supervisors, and commanders.

(6) Evidence that ninety percent of examinees completed written component within four hours, and all examinees completed hands-on component within four hours.

c. Complete SQT package.

(1) Complete administrative instructions for written and hands-on components.

(2) Sequence of tasks also standards for each step and each scorable unit in hands-on component.

(3) Scoring key for each item and standards for scorable units in written component.

10. Pilot test hands-on component to determine feasibility in FORSCOM, USAREUR, ARMY RESERVE, and NATIONAL GUARD units.

a. Develop or modify earlier instruments to record observations of hands-on component and obtain acceptability (face validity) data.

b. Coordinate with the Field Units to arrange for evaluating the hands-on component. The unit must select and train test evaluators, who are responsible for administering the hands-on tests. TRADOC Pamphlet No. 600-11 contains guidance on using performance tests, especially pages 16-19. More than fifty soldiers should be tested at a minimum of three units: one FORSCOM, one USAREUR, and one NATIONAL GUARD or ARMY RESERVE.

c. Assemble all data and forward copies to Training Management Institute, Fort Eustis.

ISSUES IN SKILL QUALIFICATION TESTING

Stephen F. Hirshfeld, Ph.D.
Army Research Institute

The Skill Qualification Test (SQT) is a critical task based and performance-oriented method of soldier evaluation. Intentions are both to measure the professional competence of a soldier and to serve the needs of personnel management. In order to achieve these goals through a performance-oriented testing program a number of issues must be considered. These issues include scoring the SQT, determining cutting scores for verification and qualification, distinguishing between testing problems in small and large population MOSs, defining the role of mandatory tasks, specifying the use of commander/supervisor certified tasks, and at a more global level, indicating the contribution of the SQT to the Enlisted Evaluation Score (EES). This paper addresses these issues individually and whenever appropriate examines the relationships among them.

Scorable Units

Skill Qualification Tests contain a written and a hands-on component. SQTs are composed of scorable units which are based on tasks and measure knowledges as they are relevant to performance on those critical job tasks. Each scorable unit is designed to identify whether or not an individual can perform a task under some set of conditions to acceptable standards. Furthermore, all scorable units are weighted equally. An extremely critical task could be weighted more heavily by including more scorable units in the SQT that are based on that task.

In the case of the written component, a scorable unit consists of a set of test items based on a task. The number of test items in a scorable unit will probably vary from scorable unit to scorable unit. In the written component one may choose to allow incorrect responses to one or more test items of the cluster forming the scorable unit and still evaluate the unit as passed. This raises a most important issue. One must establish a standard for classification as a master

---

The views expressed in this paper are those of the author and do not necessarily reflect the views of the U. S. Army or the Department of Defense.

Requests for reprints should be sent to Stephen F. Hirshfeld, U. S. Army Research Institute for the Behavioral and Social Sciences, Department of the Army, 1300 Wilson Boulevard, Arlington, Virginia 22209.

on a particular task. These standards vary from task to task. The key point is that a standard for a scorable unit is derived from the careful consideration of the task on which the unit is based and cannot be accurately determined prior to this analysis. A primary reason for the careful determination of a standard for each scorable unit is to avoid misclassifications, particularly those where a true master on a task is scored as a nonmaster. Given that we generally test only a sample of all relevant behavior, it is not obvious that a soldier answering all items correctly in a scorable unit is more likely to be a master than a soldier answering all but one item correctly. A similar option concerning standards exists for the hands-on component as well. One may allow a certain number of incorrect steps on a task and still pass the soldier on the task.

In sum, further study is required to determine methods for establishing minimal mastery criteria. Determination of standards for mastery and nonmastery classification may become easier as sufficient data become available.

Time and Resource Constraints

Another important consideration in the SQT written component is the construction of a set of scorable units that can be finished by almost all soldiers within the allotted time. If the test is a power test and not a speed test, then almost everyone should be able to attempt each scorable unit. Otherwise the criterion-referenced nature of the test is compromised. In other words, a soldier should not be classified as a master or nonmaster on a task that he did not have time to attempt. Therefore, the SQT must be constructed in such a way that a sufficient number of tasks are considered, that these tasks are adequately covered as indicated by a sufficient number of items within each scorable unit; yet, at the same time, guarantee that almost all soldiers taking the test be able to finish it within the allowed testing period.

In the case of the hands-on component an additional consideration is the practicality of resource requirements. SQTs cannot consume an excessive amount of unit resources. Implications of this are examined in a later discussion of mandatory tasks and commander/supervisor task certification.

Cutting Scores for Verification and Qualification

SQTs are composed of tasks selected from the current skill level and the next higher skill level. The SQT provides a single overall test score based on the total number of scorable units passed on the entire SQT. Test totals provide rank-order information, viz., soldiers may be ranked according to their SQT scores. Two cutting scores (decision points) are determined; a lower passing score for purposes of verification and a higher passing score for purposes of qualification.

The lower passing score is used to determine whether an individual verifies his ability to continue his present duties. If a soldier's SQT score is greater than or equal to the lower passing score, then he verifies his current skill level. The lower passing score will be set apriori. It should reflect what is viewed by expert judgment as a minimal level of acceptable performance. The higher passing score is used to identify the most qualified soldiers for promotion and other favorable personnel actions. This cutting score will also be set prior to testing. It should be set at a level that requires a soldier to pass scorable units not only at his current skill level, but also some unique to the higher skill level. The apriori lower passing score for verification and higher passing score for qualification may in reality have to be shifted after the test has been administered. The interpretation and handling of any discrepancy between the actual test performance and the expected levels of acceptable performance are major issues. This discrepancy could be an indicator of a training gap. It could also be argued that the apriori cutting scores were incorrectly determined and/or that the standards set for each scorable unit were unrealistic. Another possibility is that the test itself is in some way deficient and requires considerable revision. As with several of the issues discussed earlier, procedures for handling and evaluating such contingencies will become clearer after sufficient data have been collected and analyzed.

Mandatory Tasks

A new concept within the SQT is that of a mandatory task. A mandatory task is a task that has a direct and significant relationship to the operational effectiveness of a weapon or some other critical system. A task can be designated as mandatory only by the CG TRADOC. If a task is designated as mandatory, then it must be included in the SQT. An issue for future consideration is whether at some point not only will mandatory task inclusion be required of the SQT, but also that soldiers be required to pass the tasks. Such a requirement would place greater emphasis on the mandatory tasks and assist in focusing on the systems crucial to mission accomplishment. However, the impact of requiring soldiers to pass certain tasks would be significant. Misclassification errors must be avoided. Incorrect classification of someone as a nonmaster on a task in this situation is far more damaging to a soldier than in one where the task represents only one point out of some larger total. Another important issue that would have to be addressed is the specification of procedures for identifying and selecting such tasks. In sum, any change from the current definition of a mandatory task should be based on a careful analysis of relevant technical issues.

Commander/Supervisor Certified Tasks

Another new concept within the SQT is that of task certification. In the case where a task is both impractical to test hands-on and difficult to simulate, yet highly important, it may be administered as a supplement to the hands-on component and be commander/supervisor certified. The set of commander/supervisor certified tasks is to be viewed as a distinct component within the SQT. If resource requirements are excessive, or the soldiers in an MOS are widely dispersed and/or in remote areas, or the MOS has a very small population that makes field validation impractical; commander/supervisor certification is a reasonable solution to a complex problem. A primary concern is to identify procedures that will guarantee equity and standardization. Determination of standards for acceptable performance on a commander/supervisor certified task is a related issue that merits closer examination.

Any future decision that would require passing particular certified tasks, would require careful attention to the same technical issues raised in the earlier discussion of mandatory tasks. However, in the case of certified tasks, problems of standardization and equity would be even more difficult to solve.

Contribution of the SQT to the Enlisted Evaluation Score

At this point it is appropriate to shift from considerations within the SQT itself to those bearing on the relationship of the SQT to the remaining categories in the Enlisted Evaluation Score. The Annual Arms Qualification (AAQ) and the Army Physical Fitness Test (APFT) are both commander/supervisor certified. Therefore, they belong to the certified component of the SQT. The Enlisted Evaluation Report Weighted Average (EERWA) is also calculated in the Enlisted Evaluation Score. Current expectations are to weigh the SQT and the EERWA sixty percent - forty percent, respectively. Any future changes in the composition of the SQT, revision of SQT components, or redefinition of concepts such as mandatory and commander/supervisor certified tasks, would not only change the SQT itself, but would also effect the interrelationships among categories of task tests included in the Enlisted Evaluation Score and possibly the weighting of each factor type. Any future decision requiring that a particular category of task tests must be passed, would necessitate change in the structure of the Enlisted Evaluation Score. In sum, any major change within the SQT would probably change the determination of the Enlisted Evaluation Score as well.

Conclusion

In conclusion, a number of important issues specific to the SQT have been introduced and implications of these issues have been described.

Resolution of these issues will not only influence the SQT, but will help determine what categories of task tests are included in the Enlisted Evaluation Score and how these categories are related. Many of the issues may be clarified or even resolved as more data become available. Careful attention to the considerations presented in this paper should help insure that the Skill Qualification Tests achieve the goals for which they were created.

PRIME: Personnel Research Instrumentalities for Measurement and Evaluation

William J. Moonan, Ph.D.
Navy Personnel Research and Development Center

Introduction

The purpose of this report is to develop the framework for a structure of "assessment" and to provide procedures for the analysis of "assessment" models developed from this structure. The author feels that this endeavor is necessary because of the apparent confusion that the topic of assessment appears to have on the minds of many involved in personnel research activities. We shall proceed by examining the fundamental aspects of assessment procedures, developing assessment models, and developing computer programs for the analysis of these models.

Background

The Topic of this report is assessment for personnel R&D projects. The author assumes that the reader is in agreement with him that the framework of assessment of personnel R&D in a frightful mess and needs clarification, c.f. Anderson's (1973) Anatomy of Evaluation, which contains a very large list of articles, concepts, definitions and techniques relating to one aspect of assessment, the Page (1975). It is assumed that the author and reader are in agreement that the topic needs simplification and clarification with respect to the specification and definition of its fundamental aspects, and that the practice of personnel assessment can be improved by postulating in simple structural framework and by the development and use of formal assessment models.

The author further assumed that the reader has had a professional education with some training on measurement and statistical topics and is motivated toward improved personnel R&D and wishes to learn of a new attempt to characterize a simple and meaningful structure for the topic under discussion, and, at the same time, to learn of some new assessment models and their associated computer programs which can be used on personnel R&D work.

Types of Assessment

A. Aspects of Assessment

By the term "personnel assessment" the author refers to the "measurement" or "evaluation" of personnel. The distinction between these processes will ultimately be realized by a technical description. We are assuming also that the "object" being assessed is usually a person, since we are dealing primarily with persons in personnel research work. Our procedures are, upon occasion, also applicable for assessing other types of objects or entities such as systems, organizations, operations, programs, tasks, etc., but we will not emphasize these applications. Our purpose primarily being to assess personnel.

The substructures, assessment and evaluation, each has an enormous literature, not to be reviewed here. The subtopic "measurement" has undergone a long and steady evolution, culminating in various latent trait models and conjoint measurement procedures. The subtopic "evaluation" is more diffuse and less mathematically structured, and has a literature in the education and psychology fields primarily related to evaluations of complex programs rather than personnel. These evaluations frequently consisting of measurements or their functions. We shall not pursue this course in our development, rather we shall tend to make the process of personnel evaluation have analysis procedures similar to those of measurement.

B. Elements of Assessment

We have already identified the object or entity of assessment as one of the elements of this topic. Another element is the assessment instrument consisting of assessment items, each of which is associated with a set of item response categories. The assessment instrument in measurement situations is a test or examination whereas the instrument in evaluation situations is usually either a questionnaire, checklist or inventory. The response categories of the items of a test may be "multiple choice." For our purposes we consider that each item has only two operational responses categories, i.e., correct or incorrect. Evaluation instruments usually also have multiple category responses, which we assume are also expressed as levels of quality, several rather than two.

We postulate the existence of another assessment entity to which we append the label "assessor," whose function it is to assess the object of assessment, who is or isn't another person. In measurement the person is using his own latent abilities or competencies to designate the response categories of the items of the assessment instrument. Whereas for evaluation situations, the assessor usually utilizes his subjective judgment about himself or another entity to designate the appropriate response categories of the items of the assessment instrument.

As promised the previous two paragraphs provide our distinction between the concepts of measurement and evaluation. We summarize the various elements of assessment as follows in Exhibit 1. We realize, in truth, that our distinction between "measurement" and "evaluation" is not as sharp as we would like to make it. More realistically we may consider

EXHIBIT 1

Elements of the Assessment Process

	Assessment Type	
Assessment Entity	Measurement	Evaluation
Assessee	testee	person/entity
Assessor	testee	evaluator
Item response designator	test, examination	evaluator
Assessment instrument	problem solving,	questionnaire,
Assessment items	recognition or recall items measuring latent abilities, or items of interest and attitude	checklist statements or questions of status, opinion, value, skills and achievement
Item response categories	dichotomous; correct or incorrect	polychotomous; levels of quality, desirability or value
Essential Process	use of latent abilities	subjective judgment

"Assessment" to be a bipolar concept with some assessment situations more measurement-like than evaluation-like or vice versa. Fortunately the distinction is not critical and an appropriate assessment model for some R&D assessment project can easily be located among those we shall propose.

In addition to these factors, we consider some further elements of importance. One of these is called the Assessment Sample, which consists of those entities to be assessed once the assessment procedure has been formalized after an analysis. This analysis is usually made on data provided by an Estimation Sample from which the parameters of the assessment model are estimated and the scoring procedure formalized. The purpose of the analysis is to develop the scoring procedure to be used for providing quantified magnitudes estimating abilities or competencies in measurement situations or the "values" of the assessees in evaluation situations. Other purposes are to analyze the assessment instrument items and to examine the "fit" of the response data to the assessment model.

C. Assessment Model Spectrum

In order to characterize the assessment system proposed herein, we wish to define a set of symbols and their definitions which collectively we designate as the Assessment Model Spectrum. The purpose of this spectrum is to provide a simple structural characterization of an assessment problem and provide for the use of this spectrum for identifying the appropriate assessment model analysis. Our spectrum consists of the following nine parameters each having one or more corresponding designators as shown in Exhibit 2.

The first parameter, Aspect, indicates that the assessor is considering the assessment problem to be either a measurement problem or an evaluation problem in the sense described previously. Our assessment procedure assumes that the assessor is assessing basic characteristics or traits that are not directly observable and therefore these traits are referred to as latent traits, or he is evaluating some value. It is the purpose of the process to estimate a person's standing on a value or trait, which can in turn be used for other operational or research purposes. In measurement situations, these traits are referred to as abilities, whereas for evaluation situations, they are referred to as values, in each case being further characterized by the name of the ability or value being assessed. This name being listed as the second assessment spectrum parameter.

The parameter "instrument" refers to the type of assessment instrument utilized, most frequent types being tests, questionnaires, or checklists. The number of items on the assessment instrument is listed in the fifth parameter. Our procedures require the assessment instrument, no matter

EXHIBIT 2

Assessment Model Spectrum Parameters and Subparameters

Nr.	Parameter Name	Subparameter
1.	Aspect	M = measurement (primarily) E = evaluation (primarily)
2.	Trait	L = latent trait name to be measured V = value evaluated
3.	Entity	P = person S = system O = object or other
4.	Instrument	T = test or examination Q = questionnaire C = checklist O = other
5.	Items	I = number of assessment instrument items
6.	Responses	M = number of response categories
7.	Application Sample	N = size of application sample
8.	Estimation Sample	E = size of estimation sample
9.	Assessment Process	A = latent ability or competency S = subjective judgment

what its form, to consist of a subset of statements, questions, or whatever, which may be labeled as items. Each item is associated with a set of common response categories, the number of which we describe by the parameter M. If M = 2, the response categories are described as binary or dichotomous, where if M > 2, the response category type is called polychotomous. In multiple choice tests, M = 2, since the response is judged correct or incorrect, in spite of the fact that the testee has more than two alternatives from which to choose for each item. Our concern is for the number of "operational" response categories to be used for analysis purposes and not for the number of instrumental response categories. In our assessment model system, the response categories for evaluation instruments represent levels of quality possibly realized by the assessee relative to the items' substantive content. For example, for the item "How would you rate this instructor with respect to his general teaching ability?", for which we may have the response categories:

poor
fair
adequate
good, or
excellent

Finally, we are concerned with the existence and size of two sets of data resulting from persons for whom response information may be available. The first sample, called the Application sample is not ordinarily used for analysis. It is this sample, the members of which are to be assessed, once the assessment procedure is defined. In some cases, this application sample is not immediately available, since the assessment procedure is being developed and to be applied some time in the future. The other sample, called the Estimation sample, if it exists, provides the response data to be utilized for it estimating the parameters of the assessment model. Part of our assessment system provides models to be utilized for those occasions, usually occurring in evaluation situations, for which an Estimation sample does or cannot exist because of operational constraints. There are occasions, usually in measurement situations, when the Application and Estimation samples are identical. This circumstance causes no difficulty for the models of our system. For most measurement model analysis, we require an Estimation sample of adequate size.

The parameter "entity" is either a person, system, object or some entity which can be observed producing or made to produce information which can be transformed into assessment instrument response categories.

D. Models of Assessment

The modern scientific approach to measurement theory may be said to have begun with the publication of Lord's (1952) "A Theory of Test Scores" wherein the fundamental concept of "true scores" was postulated. A true score model may be characterized by the equation

(1) $X = T + E$ where

X represents the observed test score, T represents the true score, and E represents random error. Lord extensively published research work on this and related true score models during the score of years 1950-1970.

A related model, which for lack of a better name, we shall call the "Minnesota" model,* is

(2) $R = A + I + e$ where

R represents the response made to one of the items of a test. A represents the ability of the item respondent, I represents the item effect, and e is the random error associated with the item response. In our opinion, (2) represents the measurement model commonly assumed to exist in the minds of many personnel research workers, although its analysis is not carried forth in the popular measurement textbooks, including Gulliksen's (1950), in the fields of education and psychology. Practitioners intending to utilize the concept of (2) usually proceed to construct a measuring device whose item responses are then judged as correct or incorrect and which are accumulated into total scores. These scores are not for the purpose of estimating the ability of the testee as would be the objective if the practitioner were interested in measurement, but the scores are frequently utilized for relating them to other information and as such represent statistical, rather than measurement, objectives. As a consequence, we attach the terms "pseudo-measurement" or quasi-measurement" to this practice.

Various classes of latent-trait models are described by Lord and Novik (1968). Distinctions of these models are accomplished by denoting the parameters involved in the model specification. These parameters may be listed in Exhibit 3.

Some or all of these parameters are assembled into a mathematical expression describing the probability of a given type of $r(n,i)$ as a function of the parameters used in the model. This model is then said to be a latent trait measurement model. The so-called Birnbaum (1968) three-item parameter model utilizes the three-item parameters $\delta(i)$, $d(i)$, and $g(i)$ as well as $\alpha(n)$ in developing an expression for the probability of a correct response. The Birnbaum (1968) two-item parameter model involves the two-item parameters of $\delta(i)$ and $d(i)$, as well as $\alpha(n)$, for the same purpose.

---

*The label "Minnesota" is appropriate since Professors Palmer O. Johnson, Cyril J. Hoyt, and others used this model, during 1940-1960, as an analysis of variance model to instruct students and carry out, and develop various psychometric analyses at the University of Minnesota.

EXHIBIT 3

Parameters of Assessment Models

α (n): a real number representing the ability or latent trait of the nth person.

δ (i): a real number representing the "difficulty" of the ith item.

d (i): a real number representing the ith item's discrimination power.

r (n, i): the response made to the ith item by the nth person.

g (i): a real number representing the effect of guessing the correct answer to the ith item.

Rasch (1960) also developed a variety of latent trait models one of which, the Binary Measurement Model, BMM, develops the expression for the probability of a correct response as a function of δ (i) and α (n) only. He was able to show that certain desired measuring properties called "objectivity" resulted from this specification. Briefly, objectivity, means that we can achieve both

(a) person-free item measurement, and
(b) item-free person measurement by this specification.

The Rasch model does not actually utilize δ (i), but rather, another parameter designated as ε (i) which we may assume is conceptually and inversely related to δ (i) and referred to as "item easiness."

In our assessment system, we shall utilize the BMM as the basis for measurement and shall utilize generalizations of this model for polychotomous responses as the basis for making quantitative evaluations. We do this to maintain a degree of continuity of process, and for the reason that the Rasch models represent the simplist latent trait model and therefore can be analyzed mathematically and computationally with minimal skill and effort.

We conclude this section by indicating the assumptions required by the BMM. These are:

(1) Test items are scored dichotomously.

(2) The probability of a correct response is only a function of the ability, α (n), of the respondent and the easiness, ε (i), of the item.

(3) Responses to the items are stochastically independent given α (n) and ε (i).

The Prime System

A. Introduction to PRIME

Any assessment is the result of measuring or judging something. That is, applying some assessment process or procedure, some number or code is recorded as a result. The purpose of making an assessment is to use it for doing something. Hence the object of making assessments is to provide a basis for action.

If the assessment is related to mental measurement; the "assessment process" is usually an intellectual function, such as problem solving or the use of some memory process. In these kinds of situations, the "assessment process" is carried out by the person (entity) being measured.

If the assessment is related to evaluation, the "assessment operation" (a judgment) is often carried out by an evaluator with respect to the entity being evaluated.

In either event, the number or code resulting from assessment is a specification of a response category, i.e., a multiple choice selection or designation of a value category of assessment instrument items.

As the acronym "PRIME" indicates, it is a system devoted to "instrumentalities" for making measurements and evaluations in personnel research problems. These instrumentalities represent measurement/evaluation models and their associated computer programs designed to operate on computers available to NPRDC personnel:

The PRIME programs are enumerated as follows:

FRM: The Frequency Ratio Method (Binary Measurement Model)
TIM: The Inspection Model
PAM: The Polychotomous Assessment Model
PMM: The Polychotomous Measurement Model
HELP: The Handy, Empirical Little Prioritizer
(MAXKO): Auxiliary PRIME Program

The objective of the PRIME system is to bring "order" into the topic of assessment and to provide models, and their associated computer programs, to workers in the field of Personnel Research.

We next provide a brief overview of each instrumentality and ultimately indicate an application for it in Personnel Research Programs.

B. Briefs of PRIME Instrumentalities

FRM: The Frequency Ratio Method. This program is related to the Rasch Binary Measurement Model (BMM) discussed by Moonan (1969, 1974) and represents a new and simpler method for estimating the "easiness" of measuring instrument items. The program is offered as a possible substitute for a more extensive IBM 360/65 program called "Max" in the NPRDC Computer Program Library. FRM operates on the NPRDC GA 1830 computer system. Documentation and an example using FRM for analyzing BMM data is provided in Moonan and Covher (1975a).

FRM is useful in those measurement situations where the BMM applies and some cognitive domain characteristics, such as intelligence, ability, aptitude or competency is required to be measured by items whose responses are judged operationally to be dichotomous (e.g., correct or incorrect). BMM and FRM are applicable to most personnel measurement problems where the items scoring parameters are 0, 1.

PMM: The Polychotomous Measurement Model. Situations occur in Personnel Research, however, when it is not convenient or desirable to measure an entity property by using dichotomously scored items. Such occasions arise when measuring attitudes, interests, and motivations.

For these purposes we can use a generalization of the Rasch BMM to the polychotomous item response case Rasch (1970). We refer to this model as PMM. Anderson (1973) developed a conditional maximum likelihood computer program for this model. The author has attempted to secure this program for PRIME in the following ways:

(a) card punching a listing of the program as contained in Andersen, et al. (1973);

(b) requesting a program copy from ETS's Data Processing Division;

(c) request a program copy from Allerup, P. and Sorber, G. (1975).

Procedure (a), because of poor printing, produced a program deck that would not compile on NPRDC computers. It turned out, that in spite of Andersen et al. (1973), ETS did not have an operating program either. They offered information about a similar proprietary University of Chicago program. We have requested a PMM copy from Denmark, and have confidence in receiving it but have, as yet, not acquired it because of the delay due to distance.* Nevertheless we include PMM in the PRIME System. The chief advantage of Andersen's PMM program is that it obtains maximum likelihood estimates of the PMM parameters including the scoring parameters of the polychotomous responses to the assessment instrument items.

PMM is useful for measuring interests and attitudes as well as for analyzing evaluative instrument data.

PAM: The Polychotomous Assessment Model. Because of early concern over acquiring Andersen's PMM program, it was decided to develop an approximation program, called PAM, to the PMM program. This model handled polychotomous response data, but reated response categories independently. Rather than estimate the item scoring parameters, it defined them mathematically. The item parameter estimation procedure used was that derived for use with program FRM. Using the scoring parameters, entity scores can be derived and estimates of entity attitude, interest or whatever can be obtained.

Documentation for PAM is contained in Moonan and Covher (1975b) and provides an example using SVIB data collected from naval personnel.

TIM: The Inspection Model. Occasionally in Personnel Research, there occurs a situation wherein it is either infeasible or impossible to collect response information on the items of an assessment instrument from an Estimation sample, yet assessments must be made on an Assessment sample. If the assessments are evaluations and the items are polychotomous, the TIM model is appropriate. The item parameter information is obtained from item "importances" supplied by the evaluator.

---

*Received 31 July 1975.

Entity "scores" are obtained from a weighted average of the frequencies of item response categories (obtained during evaluation) and item response scoring parameters. A TIM model entity parameter, y, associated with each possible score is provided by the program. Such an estimate provides knowledge of the amount of the quality, measured by the assessment instrument, possessed by the evaluated entity. Please note that TIM is the first PRIME instrumentality discussed which does not require an Estimation sample which provides response data from which item parameters can be estimated. Documentation for TIM is contained in Moonan, et al. (1975c).

HELP: Handy Empirical, Little Prioritizer. In those assessment situations frequently associated with decision making, and others, one is obliged to provide quantified measures of values or utility. Often this process involves the consideration of several factors of importance. HELP calls these factors "Dimensions of Value" and requires the user to provide a consistent ratio scale of the importance (to him) of each Dimension. Each entity of the Assessment sample is then judged as to quality on each Dimension. This information is then synthesized by the HELP model into a quantified measure of entity value. These values then being used with other decision making, statistics or Operations Research models for the purposes at hand.

Such procedures, which quantify multiple-dimensional judgments, are rarely available and few are as simple as HELP. Consequently it is very important for Personnel Research purposes. The HELP documentation is given in Moonan and Covher (1975d).

MAXKO. The purpose of this program is to provide maximum likelihood estimation of the "value" parameter in the TIM and PAM model programs. MAXKO is the polychotomous generalization of subroutine MAXCO used in FRM for the dichotomous response case. The program prints a "scoring table" associating a score derived from a polychotomous assessment instrument with the estimate of the value parameter of the assessment model. The method of estimation is that of iteration on the model equation on p. 2 of Moon and Covher (1975e).

We have now listed and briefly described six PRIME programs.

For clarification purposes the following charts Exhibit 4, and Exhibit 5, characterize the relationship of 4 of the PRIME instrumentalities in Exhibit 4, and the assessment parameters of each model are shown in Exhibit 5.

EXHIBIT 4

Relations of Four PRIME Instrumentalities

Operational Response	Item Response Data	
	Objective	Subjective
Binary	FRM (BMM)	TIM
Polychotomous	PMM	TIM

EXHIBIT 5

Specification of Assessment Parameters for Five PRIME Instrumentalities

Parameter	FRM	TIM	PAM	PMM	HELP
Aspect	M	E	E,M	M	E
Trait	L	V	L,V	L	V
Entity	P	P,S,O	P	P	P,S,O
Instrument	T	Q,C	Q,C	Q,C	Q,C
Responses	M=2	M≥2	M≥2	M≥2	M≥2
Estimation Sample	Yes	No	Yes	Yes	No
Application Sample	Yes	Yes	Yes	Yes	Yes

Note. See Exhibit 2 for definitions of symbols.

V. PRIME Applications and "Measurement Fads"

Because the pursuit of good measurement and evaluation is an active one in the course of Personnel Research, we should expect that the PRIME instrumentalities should enjoy wide utility.

In the measurement field there exists the tendency toward "fads." We have already seen a long period of the fad of "Norm-referenced measurement" wherein test items and tests are calibrated relative to a sample of persons from populations of interest.

We are presently witnessing the beginning of the period of the fad of "Criterion-Referenced measurement" wherein item and test calibrations are not so important, rather standards of a priori achievement are established and persons and groups of persons are compared to these standards.

"Individual-Centered measurement" is a fad which is merely starting and whose flood we can expect in the not too distant future. Concern here is for:

(a) The fit of item responses to the measurement model used.

(b) Items whose model parameters are invariant from population to population.

(c) Person ability or competency parameters of the model to exhibit invariance with respect to the items of the collection used to measure persons.

In spite of numerous attempts by others to achieve properties (b) and (c) above only the Rasch BMM model, used in the PRIME system, has these properties and the FRM program estimates the model parameters more simply and efficiently of the several methods available to do this. Thus the BMM-FRM system becomes even more important and unique as measurement procedures for Personnel Research work.

Exhibit 6 indicates which PRIME instrumentalities are likely to be useful tools of research in each of nine research program areas at NPRDC. Also specified are those variables which are required to be measured or evaluated and those variables for which values or utilities need to be assessed.

Let us examine one area in some detail, namely program 7, Measure Job Performance. New attempts at measuring on-the-job performance have never been very successful nor frequent. Indeed this is a difficult research area and useful tools and models are rare. We do know, however, that relevant data are sometimes at hand. For example, McDowell (1972) has collected "task" data from large samples of personnel in several Navy ratings. These data provide "level of quality" responses for each of several hundred rating-related tasks and therefore could be used with PAM or PMM to obtain a measure of on-the-job competency for each questionnaire respondent. It appears that the "difficulty" in this area can, at least in part, be resolved by an effective interrelationship of the fields of Occupational Analysis and psychometrics. Each field seems to have already developed the appropriate technology to make the key contributions to solve the problem. The critical point being to use the categorical response data resulting from an occupational task survey together with an assessment model which can accommodate such data and subsequently derive a measure of person competency. These concepts have utility beyond research at NPRDC since Task Inventory data collection and analyses are being carried on elsewhere in the Navy (c.f. NOTAP) and by other services.

The assessment of "Effectiveness" of Persons, Systems or Objects seems of general importance. To use PRIME for this problem it appears that we must take the "bull by the horns" and indicate, for each application, the issues to be considered important for assessing effectiveness as well as the relative importance of each issue. The final say in these matters is of course the perogative of high level policy makers, but, at the research level, we can now provide a procedure and first attempt approximations to "effectiveness" policy. PRIME instrumentalities TIM and HELP seem to be appropriately relevant models, among others, for assessing effectiveness. We conclude this section by summarizing applications in Exhibit 6.

Summary

Our problem is called "Assessment" and our purposes were to develop a structure by which assessment models could be uniquely characterized and to develop assessment models and their associated computer programs which could be used for personnel research purposes.

To these ends two aspects of assessment, namely measurement and evaluation were identified. Their distinction lying mainly in the process used to respond to the items of an assessment instrument. There are, however, often elements, or their combinations, which can be used to clarify the distinction.

The parameters of the assessment models were identified. These were variously combined and operationalized into either measurement or evaluation models. The intention of this process being to make evaluation more measurement-like by using and developing evaluation models, for polychotomous response information, which are patterned after Rasch latent-trait measurement models. Also developed were some special models for use in

EXHIBIT 6

Potentially Relevant PRIME Instrumentalities to NPRDC Program Areas

No.	Program Title	Important Assessment Variables	Important Values or Utilities	Potentially Relevant PRIME Instrumentalities for Program Assessments				
				FRM	PAM	PMM	TIM	HELP
1.	Acquisition and Initial Service	Performance	Recruit Assignments	*	*	*	*	*
2.	Career and Occupational Design	Job Performance Occupational Success		*	*	*	*	*
3.	Attitude and Motivational Research	Attitudes, Motivation Operational Readiness			*	*	*	*
4.	Personnel Effectiveness	Personnel Effectiveness			*	*	*	*
5.	Training Technology	General Evaluations					*	*
6.	Test Training Systems	Training Systems		*			*	*
7.	Measure Job Performance	Job Performance		*	*	*		
8.	Human Performance		Human Effectiveness		*	*	*	*
9.	People and Organizational Management	Organizational Effectiveness, Motivation, Processes			*	*	*	*
10.	Management Systems		Organizational Structure			*	*	*
11.	Human Information Processing		Decision Making Methods and Values			*	*	*

decision making paradigms and for assessment situations for which an Estimation sample does or does not exist.

Analysis of these models and their associated data is provided for by a system of computer programs called PRIME. In all, five programs were developed, illustrated and documented. Four are now operating on the GA 1830 computer system at NPRDC in San Diego, California.

Applications of these models and computer programs to personnel research areas of the NPRDC research program was discussed.

VII. REFERENCES

Allerup, P., & Sorber, G. (1975). "The Rasch Model for Questionnaires." Report of the Danish Institute for Educational Research, Copenhagen, Denmark.

Andersen, E. B. (1972). "A Computer Program for Solving a Set of Conditional Maximum Likelihood Equations Arising in the Rasch Model for Questionnaires," (unpublished). Educational Testing Service, RM-72-6, Princeton, New Jersey.

Andersen, E. B. (1973). "Conditional Inference for Multiple-Choice Questionnaires," British Journal of Mathematics, Statistics and Psychology, 26:31-44.

Anderson, S. B. Ball, S., Murphy, R. T., & Rosenthal, E. J. (1973). "Anatomy of Evaluation," Educational Testing Service, Princeton, N.J.

Birnbaum, A. (1968). Some latent trait models and their use in influencing an examinee's ability. Part V of F. M. Lord and M. R. Novick, Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

Gulliksen, H. (1950). Theory of Mental Tests. New York: John Wiley.

Lord, F. M. (1952). "A Theory of Test Scores." Psychometric Monograph, No. 7.

Lord, F. M., & Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley Publishing Company.

McDowell, H. C. (1972). "New Concepts to Improve Training and Shipboard Manpower Utilization in the Engineering and Hull Ratings: Statistical Analysis." Navy Personnel Research and Development Center, San Diego, California, NPTRL SRR 72-26, June 1972.

Moonan, W. J. (1974). "COMET: Contributions to Objective Measurement and Evaluation of Trainee Competency," Proceedings of the 1974 MTA, Oklahoma City, Oklahoma.

Moonan, W. J. (1974). "CHAROSEL: A Computer Program for Qualitative Criterion Prediction Problems," Proceedings of the 1973 MTA, San Antonio, Texas.

Moonan, W. J., & Covher, M. (1975a). "FRM," Tools for Analysis and Decision, FACT DOCUMENT No. 32, Naval Personnel Research and Development Center, San Diego, California.

Moonan, W. J., & Covher, M. (1975b). "PAM," Tools for Analysis and Decision, FACT DOCUMENT No. 34, Naval Personnel Research and Development Center, San Diego, California.

Moonan, W. J., Southwick, P. C., & Covher, M. H. (1975c). "TIM," Tools for Analysis and Decision, FACT DOCUMENT No. 33, Navy Personnel Research and Development Center, San Diego, California.

Moonan, W. J., & Covher, M. H. (1975d). "HELP," Tools for Analysis and Decision, FACT DOCUMENT No. 37, Navy Personnel Research and Development Center, San Diego, California.

Moonan, W. J., & Covher, M. H. (1975e). "MAXCO," Tools for Analysis and Decision, FACT DOCUMENT No. 35, Navy Personnel Research and Development Center, San Diego, California.

Page, E. B. (1975). "Evaluating 'Evaluation'," The Review of Education, 1(1), 133-140.

Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Danish Institute for Educational Research.

Rasch, G. (1970). "An Individual-Centered Approach to Item Analyses with Two Categories of Answers." Unpublished, 1969 AERA Conference on Rasch Models, Los Angeles, California.

VARSEL: VARIABLE SELECTION FOR MULTIPLE-PURPOSE PREDICTION SYSTEMS WITHOUT USING EXTERNAL CRITERIA

By
R. Bruce Gould
Raymond E. Christal
Occupational and Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland Air Force Base, Texas

I. INTRODUCTION

Test and inventory developers frequently obtain large item pools from which to select a final set of items. Laborious efforts to define the area of interest and to develop corresponding items culminate in the final critical developmental stage--selection of the set of items to be included in the operational instrument. The usual item selection approach is to administer the entire pool of items to a sample population, correlate the items with some criterion, and select those items having the highest correlations with the criterion. Frequently, suitable external criterion measures do not exist and in some cases the criteria have not yet been adequately defined. The selection procedures and criteria problem is similar for developing test batteries except that the variables of interest are complete tests instead of specific items.

In the absence of suitable external criteria, the standard approach to selecting the final set of items is to factor analyze the item pool and use factor scores as the criteria for the item selection phase. Use of factor scores has three primary disadvantages. First, selecting items which correlate with factor scores tends to produce a set of highly related or redundant items with a corresponding loss of items accounting for relatively unique components of score or item specific variance. It would not be at all unusual for factor scores to account for little more than half the variance represented by an intercorrelation matrix and thus substantial variance from an original set of variables could be lost by selecting a subset of variables which only reproduced the factor scores. Secondly, many item and test scores fail to meet the requirements for valid factor analyses as explained by Guilford (1952). Linear dependencies among scores, skewed variable distributions, and factor extraction inadequacies are only a few of the data and procedural discrepancies which can introduce criterion error when factor scores are used in item selection exercises. Finally, use of factor score criteria unnecessarily limits the future applications of many measurement instruments, particularly diagnostic instruments, because of the specificity of the item selection criteria.

Some investigators have recognized the tendency for correlational techniques to select redundant items and proposed accretion item selection procedures which take into account item intercorrelations during the selection process (Bryson, 1972; Moonan and Pooch, 1966). Their accretion procedures partially solve the redundancy problem but have limited effect on the loss of item specific variance and subsequent restrictions on the future utility of the measurement instrument. The problem remains--how to start with a large pool of items or tests and select the smallest subset of variables which can capture the domain of non-error variance measured by the full set of items or tests?

This paper proposes an item selection algorithm which test developers may find useful in constructing multiple-purpose measurement instruments. The algorithm is titled VARSEL, variable selection for multiple-purpose prediction systems without using external criteria. A description of the VARSEL algorithm, including theoretical considerations and computational options, will be followed by two example applications; (1) selection of a subset of job attitude items from a pool of 348 items and (2) selection of a subset of tests from a battery of 56 tests.

II. VARSEL

Purpose

VARSEL was originally developed to meet an operational requirement and the purpose of VARSEL will be explained in the context of that operational problem. An experimental Occupational Attitude Inventory (OAI) was being developed to measure all facets of job satisfaction in the military work environment (Tuttle, Gould, and Hazel, 1975). Planned usage of the OAI was as an Air Force wide, a career-ladder specific, or a within career ladder job satisfaction diagnostic instrument as well as for predictive studies of reenlistment, performance, and attitude changes. These multiple objectives for the OAI precluded the use of any particular external criteria in the item selection process. During the developmental stage, a comprehensive pool of 348 items was developed to sample the domain of job attitudes of Air Force enlisted airmen. The problem was, how to select the smallest subset of these items which would capture the domain of attitude variance measured by the full set of 348 attitude items? Therefore, VARSEL was developed to identify a minimum set of selected items (predictors) which can account for the variance of the set of nonselected items.

Developmental Considerations

Two systems, a deletion and an accretion system, using multiple linear regression techniques, were considered for the item selection process. The deletion system would have sequentially removed items which were most highly predictable from remaining items in the pool. However, it became apparent that the approach could discard an item which had small amounts of variance uniquely held in common with several items and subsequently delete some of those common variance items. The resulting loss of unique variance could defeat the purpose of the selection system. Conversely, an accretion system does not present this problem, and so it was used for the VARSEL algorithm.

Figure 1. Accretion Item Selection Technique

At each iteration, VARSEL . . .

- Uses a multiple linear regression routine to predict each unselected item where selected items (S) are used as predictors. Results in N-S R^2 values. N=total number of items in input correlation matrix.
- Subtracts resulting R^2 values from "1.0"; $1.0-R^2$=amount of unpredicted item variance. Reliability (r_{11}) used instead of 1.0 if r_{11} is known; $r_{11}-R^2$=amount of unpredicted reliable variance.
- Selects item which has greatest amount of unpredicted variance; i.e., selects item which contributes greatest amount of unique variance (unique reliable variance if r_{11} is used) to the prediction system.

The accretion or accumulative system (summarized in Figure 1) was designed to start with a single item or set of items and iteratively select additional items which would contribute the greatest amount of unique variance to the prediction system. At each iteration the selected set of items (S) was used to predict (R^2) the remaining N-S individual items. The resulting N-S R^2 values were then individually subtracted from "1.0". The $1-R^2$ values represented the amount of item variance unpredicted by the set of S items and thus the amount of unique variance the item would contribute to the prediction system. The item with the highest $1-R^2$ value was selected for inclusion in the S set of items, and the process continued until any one of several cutoff parameters was reached.

As thus far expressed, there was a major weakness in the selection system. Unique or specific item variance is composed of true and error variance. Items with large error variances would likely be selected even if their true variance was "perfectly" predicted by previously selected items. Therefore, the selection system was designed to take into account item reliabilities, when they are available, by subtracting the R^2 values from the respective reliabilities rather than using the $1.0-R^2$ item selection criteria. This can be done because a test's reliability r_{11} is a measure of reproducible test score variance. Items thus identified were selected according to their contribution of unique "reliable" variance to the prediction system. Use of reliability values in the selection process is particularly useful when tests or subtests are identified for inclusion in a test battery.

Figure 2. VARSEL Options

Cutoff Options -- Stop when ...

- A predetermined number of items (S) are selected; or
- Prediction system unique variance contribution of the candidate selection item falls below a specified minimum; or
- All items have been selected.

Selection of Starting Variable(s)

- VARSEL selects most unique of N items based on unique reliable variance if r_{11} values are available; or
- User selects starting item or set of items

Iteration Print Suppression Options also Available

Several VARSEL iteration cutoff options may be specified by users as shown in Figure 2. Variable selection iterations may be stopped when (1) a predetermined number of variables has been selected; (2) prediction system contributions of unique reliable variance of the next item to be selected falls below some specified minimum; or (3) all items have been selected. Options are also available for selecting the initial or starting variables as follows: (1) VARSEL will select the most unique of N items on the first selection iteration; or (2) the user may specify the first item selected; or (3) the user may specify a group of items to be initially selected. The third option permits the developer to select variables he definitely wants in his operational inventory. For example, those interested in ensuring that a previously defined factor structure of a pool of variables is maintained can select for each factor a few variables having high factor loadings as the initial system (S) predictors.

Figure 3. Selected System Requirements and Computing Efficiency[1]

Required Data

- UPPER triangular correlation matrix; $N \leq 400$
- Reliabilities (1's are default values)
- Selection cutoff and print options (default is 10 items and print all iterations)
- Initial selected item IDs (default is most unique of N items)

System Requirements

- 80K core memory if N=400
- Tape drives, disk packs, or drum memory
- FORTRAN V compiler

Efficiency Example

- 54,285 R^2 values computed by Cholesky's routine in 7.75 minutes or .0086 second per R^2
- Complete VARSEL processing of 330 variable matrix in 8.2 minutes including input and output when starting variable is given

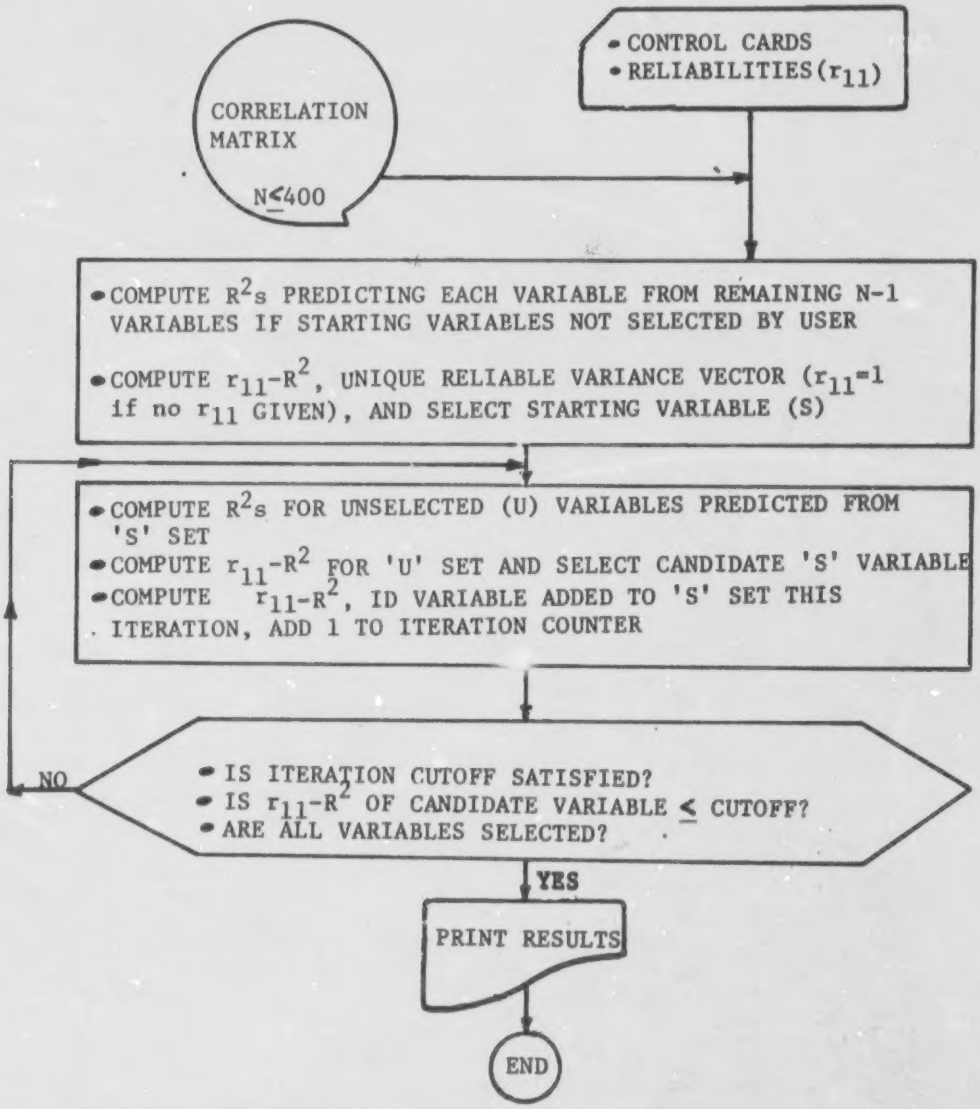
---

[1] FORTRAN V source program was written by A1C David Westphal of the AFHRL Computational Sciences Division.

Operating Requirements

VARSEL requires as input an upper triangular correlation matrix, in column or row sequence. Figure 3 summarizes some of the VARSEL input and computing requirements. The maximum number of variables that can be handled is 400. Reliabilities are optional as are methods for selecting initial variables and iteration cutoff criteria. When options are not specified, the following default values are used: Reliabilities = 1; selection of most unique item

Figure 4. VARSEL Abreviated Flow Chart


CORRELATION MATRIX
$N \leq 400$
• CONTROL CARDS
• RELIABILITIES(r_{11})
• COMPUTE R^2s PREDICTING EACH VARIABLE FROM REMAINING N-1 VARIABLES IF STARTING VARIABLES NOT SELECTED BY USER
• COMPUTE $r_{11}-R^2$, UNIQUE RELIABLE VARIANCE VECTOR ($r_{11}=1$ if no r_{11} GIVEN), AND SELECT STARTING VARIABLE (S)
• COMPUTE R^2s FOR UNSELECTED (U) VARIABLES PREDICTED FROM 'S' SET
• COMPUTE $r_{11}-R^2$ FOR 'U' SET AND SELECT CANDIDATE 'S' VARIABLE
• COMPUTE $r_{11}-R^2$, ID VARIABLE ADDED TO 'S' SET THIS ITERATION, ADD 1 TO ITERATION COUNTER
NO
• IS ITERATION CUTOFF SATISFIED?
• IS $r_{11}-R^2$ OF CANDIDATE VARIABLE $\leq$ CUTOFF?
• ARE ALL VARIABLES SELECTED?
YES
PRINT RESULTS
END

as the starting S variable; and iteration cutoff = 10 items selected. For those interested in computer requirements, selecting from a 400 variable correlation matrix requires about an 80K core memory. Cholesky's routine (Overall and Klett, 1972; System 1360, 1966) is used to develop the R^2 values and the efficiency of the method is illustrated by the computation of 54,285 R^2 values in 7.75 minutes (.0086 seconds per R^2 value) on a UNIVAC 1108. VARSEL took 8.2 minutes to completely process a 330 variable matrix including input and output when the starting variable was given by the user. Processing required 14.7 minutes when VARSEL selected the first item. Additional time was required to calculate the inverse of the 330 x 330 correlation matrix and re-read the original matrix when the most unique item was identified by VARSEL. Cholesky's routine provides unique solution R^2 values and is usually associated with factor analysis programs. The VARSEL source program is written in FORTRAN V and the control cards are free format. A brief flow chart of VARSEL is shown in Figure 4.

III. EXAMPLE APPLICATIONS

Two example applications are given. The first demonstrates the selection of a minimum subset of items from a large pool of items and the second illustrates the selection of a subset of tests for inclusion in a test battery. The item selection example also demonstrates how to ensure that the process maintains the factor structure of an original pool of items. The test selection example demonstrates the use of reliabilities. Both examples show the VARSEL effectiveness in reducing a set of variables with little loss in system variance.

Item Selection Example

The 348 item OAI was administered to a random sample of 3,100 airmen. Previous studies have indicated that most job dissatisfaction is among first-term airmen (Gould, 1972). Therefore, responses of 1,340 first-term airmen were used in the item reduction exercise. The 348 OAI items were first factor analyzed and 36 dimensions or facets of job satisfaction identified before VARSEL processing began. Two items having high factor loadings, yet which subjectively measured different specific elements of the satisfaction dimension, were selected per factor for use as the starting set of selected items. This was done to ensure that the factor structure of the entire domain of 348 job attitude items was maintained during the selection process. Because of the extreme length and varied content of the experimental OAI, it was not practical to obtain item reliabilities so a simple estimate procedure was used. One item, "The chance to work at your own pace," was placed in the OAI twice as items number 235 and 306. The item is the same format as the other 346 items and is considered to be a relatively simple and straight forward statement of an aspect of a job. Therefore, the correlation between items 235 and 306, $r = .77$, was accepted as the best available estimate of the individual items' reliabilities (r_{11}). Instead of subtracting the R^2 values from "1.0", $r_{11}-R^2$ or $.77-R^2$ values were used to determine the potential unique reliable variance contribution of each selection candidate item. Because the same reliability is applied to all items, the r_{11} value plays little role in the actual selection process but does help to determine when to stop selecting items.

Table 1. VARSEL Iteration Summary Print Extract for 348 OAI Items

Iteration Number	1	2	3	4	5
Sum of Unique Reliable Variance	48.25	47.33	46.39	45.42	44.52
Variable Selected This Iteration	-	91	55	115	16
Selection Candidate For Next Iteration	91	55	115	16	56
Number of Predictor System Variables	71	72	73	74	75

Variable ID	And Description	$r_{11}-R^2$	$r_{11}-R^2$	$r_{11}-R^2$	$r_{11}-R^2$	$r_{11}-R^2$
1	Chance to feel you are accomplishing something	.109	.109	.107	.107	.106
2	Air Force's efforts to remove irritants	.376	.376	.373	.372	.371
3	Geographical area to which you are assigned	****	****	****	****	****
6	Need for frequent retraining within your specialty	.455	.453	.448	.448	****
91	Cost of living in the area to which you are assigned	.575	****	****	****	****
115	Adequacy of lighting in the immediate work area	.455	.454	.454	****	****
148	Opportunity for personal growth	.022	.021	.021	.021	.021
235	Chance to work at your own pace	.211	.209	.208	.208	.208

Table 1 presents an extract of the VARSEL iteration summary output obtained for the first five OAI item selection iterations. Seventy-one "user selected" variables were placed in the predictor set on the first iteration. The $r_{11}-R^2$ values for each nonselected variable is shown in the iteration column. Asterisks in the iteration column indicate that the item has been selected in the current or a previous iteration. Item 91 has the highest column value after iteration 1 so will contribute the most unique reliable variance to the predictor set and is selected to be placed in the S set of items in iteration 2.

The sum of the $r_{11}-R^2$ values is shown immediately under the iteration number. Users may find the sum helpful in determining how many items to actually select. The sum is an error sum of squares value and represents the amount of reliable domain variance that will be lost if the selection process stops at the respective iteration. In the current example, the user had no required number of items to select so VARSEL was allowed to continue until all items were selected. The user then selected the iteration which would define the final set of OAI items.

The unique reliable variance sum was used as one indicator of when to stop selecting items. Iteration 128 was selected as the stopping point resulting in a final set of 200 items for the OAI. The sum of the unique reliable variance in the set of remaining 148 items was .36 so the average loss of unique reliable variance per item was .0024. The candidate item for selection in iteration 129 had an $r_{11}-R^2 = .075$ so less than 8 percent of its true variance was not accounted for by the set of S variables. Most of the unselected items had less than 1 percent of their true variance unaccounted for by the S set of items; a highly successful result. The 200 selected items were refactored and the original 36 dimensions were obtained. The factor structure of the domain of attitude items was maintained during the item selection process.

Test Selection Example

The Army Air Force Aviation Psychology Program Research Reports (Guilford and Lacey, 1947, p. 903) contain an intercorrelation matrix of 65 experimental and operational classification (aptitude) tests developed and used during World War II with pilots and navigators. The tests had reliability coefficients available and are here used as a test of the VARSEL algorithm for reducing a test battery to a subset of tests measuring the same domain of reliable variance. Reliabilities ranged from .49 to .94 but the mean reliability was .80.

The test battery had undergone extensive and laborious research efforts to develop unique tests. The intercorrelation matrix had r's ranging from -.01 to .62 but most of the r's were in the teens and low twenties. These r's are remarkably low for a large number of aptitude tests. The test battery already contains seemingly unique tests and should pose a rigorous trial for the selection routine. Before VARSEL application, nine tests were removed from the matrix because of obvious inconsistencies in the correlation matrix.

Table 2. VARSEL Iteration Summary Print Extract for 56 Tests

Iteration Number	1	10	20	30	40
Sum of Unique Reliable Variance	25.3	17.34	10.33	5.30	1.83
Variable Selected This Iteration	-	9	17	3	18
Selection Candidate For Next Iteration	53	56	44	34	14
Number of Predictor System Variables	8	17	27	37	47
Variable ID And Description	$r_{11}-R^2$	$r_{11}-R^2$	$r_{11}-R^2$	$r_{11}-R^2$	$r_{11}-R^2$
1 Map Memory	.395	.340	.286	.244	.203
4 Estimation of Length	.419	.359	.335	.269	.239
6 Directional Orientation	.451	.398	.326	.249	.209
14 Objectivity of Perception	.507	.372	.343	.308	.250
16 Aerial Orientation	.408	.339	.249	.219	.193
24 Pattern Assembly	.432	.357	.329	.265	.239
34 Tool Function	.422	.400	.383	.354	****
35 AAF Word Knowledge	.439	.398	.285	.232	.199
39 Spatial Orientation I	.348	.313	.289	.216	.194
44 Speed of Identification	.537	.506	.471	****	****
45 Numerical Operations I	.153	.146	.138	.125	.101
53 Finger Dexterity	.759	****	****	****	****
56 Discrimination Reaction Time	.606	.569	****	****	****

Using a stop criterion of $r_{11}-R^2 = .25$ and a starting set (S) of eight tests, the iteration process stopped at the 40th iteration. There were nine tests unselected and they had 75 percent or more of their unique reliable variance accounted for by the 47 selected tests. Table 2 presents an extract of selected iterations and items from the VARSEL summary output.

Test number 14, Objectivity of Perception, has 51 percent of its reliable variance unaccounted for by the initial selected set of eight tests. Fifteen percent of test 14's total score variance, .66 (r_{11}) minus .51, is variance held in common with the eight selected tests. The unaccounted for reliable variance decreases to 25 percent by the 40th iteration when 47 tests have been selected. By definition the test with the largest column value is the candidate item for the next iteration, so test 14 is the candidate test for selection in iteration 41. Since item 14's contribution of unique reliable variance is equal to the $r_{11}-R^2$ cutoff, the iteration process stops and nine tests are unselected. The nine tests may be removed from the test battery as having most of their variance in common with the selected set of tests.

Removing 16 percent of the tests from a test battery which appeared to be composed of unique tests could represent a considerable savings in personnel, testee and testor, time. A comparison of the reliabilities of the unselected tests and the correlations of those tests' scores with scores predicted from the selected tests may clarify the efficiency of the VARSEL process. The values are in Table 3.

Table 3. Unpredicted Reliable Variance, Reliabilities, and Predictability Of Test Scores For Nine Unselected Tests

Variable ID	Description	Unpredicted Reliable Variance	r_{11}	R
1	Map Memory	.203	.66[1]	.68[2]
4	Estimation of Length	.239	.60	.60
6	Directional Orientation	.209	.74	.73
14	Objectivity of Perception	.250	.66	.64
16	Aerial Orientation	.193	.69	.71
24	Pattern Assembly	.239	.59	.59
35	AAF Word Knowledge	.199	.79	.77
39	Spatial Orientation I	.194	.63	.66
45	Numberical Operations I	.101	.64	.73

[1] Test reliabilities as determined by split-half, test-retest, or internal consistency (KR_{20}) methods.

[2] Multiple correlations predicting obtained scores for unselected tests using a combination of test scores from selected tests.

After taking into account test reliabilities, the final unselected tests had reliabilities much below average. In a trial VARSEL run which did not use the reliabilities, several of these less reliable tests were selected early in the iteration process demonstrating the importance of using reliabilities.

Using the selected tests to predict individuals' scores on the nine unselected tests, correlations (R) between actual and predicted scores ranged from .59 to .77. In each case, the correlation between predicted and actual scores approached or exceeded the tests' reliabilities demonstrating the efficiency of the selection process.[1]

Additional Applications

Innovative users may find several additional uses for the VARSEL program. For example, most researchers who use multiple linear regression techniques will recognize the problem of overfitting the data. If the number of predictors is large and the N is small, VARSEL can be used to identify the most parsimonious set of predictors before hypothesis testing or computation of predicted scores begins. In another type of regression analysis where significant interactions are sought such as one by Automatic Interaction Detection (AID) programs (Koplyay, Gott, and Elton, 1973), the programs frequently generate multiple interactions which defy verbal descriptions. Instead of using AID approaches, investigators could generate a large group ($N \leq 400$) of potentially meaningful interactions and use VARSEL to identify the smallest significant subset of variables, a minimum subset of meaningful interactions.

IV. CONCLUSION

In summary, a computing algorithm has been developed to select minimum subsets of items or tests which capture the domain of reliable variance measured by a large pool of items or tests. Innovators may discover several additional uses for the VARSEL algorithm.

---

[1] Tests may correlate with other variables more than they correlate with themselves; i.e., correlations of test scores with other variables may exceed the test reliabilities. Maximum possible correlation of a test with another variable is equal to the correlation of the test's true and obtained scores; the value is estimated by taking the square root of the reliability coefficient.

Using an accretion method, the variable which will contribute the most unique variance if included in the final instrument, or which represents the greatest loss in domain variance if excluded, is selected at each iteration. Cutoff parameters stop the process when a specific number of items are selected or when the amount of unique variance contributed by a selected item falls below a specified value.

In demonstrations of the efficiency of the algorithm, a subset of 200 items was selected from a pool of 348 job satisfaction items. Of the items eliminated from the pool, an average of only two-tenths of one percent of reliable variance per item could not be accounted for by the selected items. In another application selecting tests from a battery of 56 seemingly unique tests, nine tests were eliminated from the battery. Using the selected tests to predict unselected test scores resulted in higher predicted vs actual score correlations than the individual tests correlated with themselves (reliabilities), thus demonstrating the efficiency of the selection process.

REFERENCES

Bryson, R. *Shortening tests: Effects of method used, length, and internal consistency on correlation with total score.* In Proceedings of the 80th Annual Convention of the American Psychological Association, Honolulu, 1972, 7,7-8.

Gould, R.B. *Reported job interest and perceived utilization of talents and training by airmen in 97 career ladders.* AFHRL-TR-72-7, AD-745 099. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, January 1972.

Guilford, J.P. & Lacey, J.I. *Printed Classification Tests.* Army Air Force Aviation Psychology Program Research Reports, Report No. 5, 1947.

Koplyay, J.B., Gott, C.D., & Elton, J.H. *Automatic Interaction Detector - Version 4 (AID-4).* AFHRL-TR-73-17, AD-773 803. Lackland AFB, Tex.: Personnel Research Division, Air Force Human Resources Laboratory, October 1973.

Moonan, W.J. & Pooch, U.W. *SEQUIN: A computerized item selection procedure.* USNPRA Research Memorandum SRM 67-8. San Diego, Calif.: U.S. Naval Personnel Research Activity, 1966.

Overall, J.E. & Klett, C.J. *Applied multivariate analysis.* McGraw Hill: New York, 1972.

System 1360 Scientific Subroutine Package. (360A-CM-03X) Version III Programmers Manual. International Business Machines Corporation: White Plains, N.Y., 1966.

Tuttle, T.C., Gould, R.B., & Hazel, J.T. *Dimensions of job satisfaction: Initial development of the Air Force occupational attitude inventory.* AFHRL-TR-75-1, AD- . Lackland AFB, Tex.: Occupational Research Division, Air Force Human Resources Laboratory, June 1975.

Capturing Performance Rating Policies From Multitrait-Multirater Data

by

Jimmy L. Mitchell, Major, USAF
Department of Psychological Sciences
Purdue University

Introduction

Users of the multitrait-multirater approach of Lawler (1967) have been relatively unsuccessful in demonstrating significant discriminant validity for their constructs. Examination of studies by Heneman (1974) and Lawler (1967) suggests that the comparison of a global performance rating with functional performance variables in the assessment of discriminant validity may be inappropriate and this may be one cause of the difficulty. However, such overall performance ratings can be very useful in supplemental analysis of multitrait-multirater data when used as a criterion for policy capturing the strategies of the various classes of raters. This procedure permits comparison of the relative weights of variables by type of rater as well as examination of the R^2 values from the prediction of overall performance. This adds considerable information to that currently developed in the multitrait-multirater approach. In this study, the rating strategies of the various types of raters in reports by Heneman and Lawler will be examined. We shall see that there are substantial differences in the percentage of overall performance accounted for and in the rating strategies of self-, peer, and supervisory raters. Some of the implications of these findings will be discussed and suggestions made for future performance criteria research.

Background

In 1967, Lawler reported a modification of the multitrait-multimethod matrix approach of Campbell and Fiske (1959) which applied the system to the issue of performance ratings. Performance dimensions were considered by Lawler to be analogous to traits and the different types of raters (e.g. superiors, peers, and self) were seen as methods. The validity of the rating scales for the various performance dimensions were evaluated in terms of the degree to which Campbell and Fiske's criteria of convergent and discriminant validity were met.

Since that time, there have been a number of studies which used the Lawler approach to deal with performance ratings gathered from more than one set of raters (Arvey & Hoyle 1974; Campbell, Dunnette, Arvey, and Hellervik 1973; Heneman 1974; Kavanagh, MacKinney & Wolins 1971; Klimoski

& London 1974; Williams & Seler 1973; Zedeck, Imparato, Krausz & Oleno 1974). Most of these studies demonstrated some degree of convergent validity but they have been generally unsuccessful in showing discriminant validity. Typically, researchers have simply applied the criteria and then have bemoaned the degree to which the data were found lacking.

Recently, some have begun to question the extent to which the convergent and discriminant validity criteria are appropriate for judging the validity of performance ratings obtained from more than one reference group (Borman 1974; Klimoski & London 1974). Borman has argued that the various types of raters have qualitatively different contact with the ratee and thus that there is differential opportunity to observe various dimensions of his performance. For example, supervisors may be very capable of judging the extent to which a subordinate "meets deadlines," but may not be able to properly judge the extent to which the subordinate "gets along well with peers." Peers, on the other hand, could validly rate the latter dimension but not the former. Therefore, according to Borman, a lack of convergent validity in each of these dimensions would reflect valid differences in observation but not necessarily a lack of construct validity.

For some uses of the multitrait-multirater approach, it is possible to argue that the discriminant validity criterion is also invalid when applied to performance rating scales. The special case in which this criticism holds is when overall performance ratings are included with the functional performance dimensions. Heneman (1974) studied self- and superior ratings using eight functional dimensions of managerial performance developed by Mahoney, Jerdee, and Carrol (1963, 1964), plus an overall performance rating. Partially as a result of the inclusion of this overall rating, within each subset no discriminant validity could be demonstrated; all eight performance dimensions correlated highly with the off diagonal row or column representing the overall performance ratings.

Although one may, as Heneman did, discuss the lack of discriminant validity as a problem of possible halo effects, inflation of ratings, and other contaminating factors, it seems more reasonable to question the utility of the multitrait-multirater approach for assessing validity of performance ratings when one includes an overall performance measure. It seems much more defensible to argue that one should not expect, in fact, should not desire, discriminant validity in this case. If a global performance rating really does measure all dimensions of performance (that is, if it is a composite or summation of performance), then we would expect each performance dimension to correlate with it. Consequently, the high correlation which precludes demonstration of discriminant validity in the multitrait-multirater matrix is actually desirable and is a necessary (but not sufficient) condition for showing the validity of the overall rating.

In spite of the fact that an overall performance rating makes the demonstration of discriminant validity inappropriate in a multitrait-multirater matrix, such data to provide an interesting way to further assess the validity and utility of performance ratings. It was noted earlier that such overall ratings could be considered as a weighted composite of the other performance dimensions. This implies that one additional way to look at the data would be to predict the overall performance variable using the performance dimensions as predictors. This would essentially be a policy capturing of the self- and the superior's rating strategies. Such a policy capturing approach has been used previously by Naylor and Wherry (1964) in a study of the rating policies of supervisors using simulated job incumbents. Their purpose was to establish if there were differences in policy between different types of jobs and various levels of supervision. They found substantial differences in rating policy among the subgroups. Their work implies that, in examining Heneman's data, we might expect to find differences in rating strategy between self- and supervisor raters since they are at different organizational levels and thus have different job perspectives. It is not possible with Heneman's data to examine the differences in rating strategy which might be the result of different kinds of jobs, although Naylor and Wherry's study suggests that such differences may also be a significant source of systematic variance in ratings.

Method

Rating policies can be expressed in terms of standardized regression coefficients (b*). These effective beta weights are calculated for each class of rater in standardized form. Since the ratings are on the same subjects, the use of standardized regression weights facilitates comparison of the rating policies across classes of raters.

To capture the rating policies from published multitrait-multirater matrices, the heterotrait-monorater triangles of the matrix are used to generate tables of intercorrelations for each class of rater. Within the heterotrait-monorater triangles, the vector of correlations between the overall performance rating and the other performance dimensions is used as a vector of validity coefficients. The heterotrait-monorater triangle is reflected and unities are entered in the diagonal so that it becomes an augmented intercorrelation matrix having the criterion (overall performance) in the last row and column.

The standardized regression weights can be calculated by the application of the SWP algorithm (Winer 1971: 125-126) to the augmented intercorrelation matrix. This algorithm is essentially a method for computing the inverse of the matrix. Once the inverse of the intercorrelation matrix is obtained, certain of the elements of the inverse matrix are used to calculate the standardized regression weights. An additional value of the method is that the reciprocal of the last element in the diagonal of the inverse represents the residual of unpredicted variance. When the reciprocal of this element is subtracted from unity, the resulting value is R^2 which represents the percentage of criterion variance accounted for by the predictor dimensions.

The R^2 can also be computed by summing the product of each of the standardized regression weights and its corresponding validity coefficient ($b^*_i r_{yi}$). This second computation of the R^2 serves to check the correctness of the calculated standardized regression weights. Additionally, the determination of this product ($b^*_i r_{yi}$) can also be useful as an estimate of the relative contribution of each dimension or variable to the obtained R^2.

Results

Using the procedure outlined above, the rating policies of the self- and superior raters in Heneman's study were calculated from the data in his multitrait-multirater matrix (Heneman 1974: Table 2). The results of these calculations are shown in Table 1.

Table 1

Relative Prediction of Performance by Self and Superiors

Dimension	Self			Superior		
	b*	r	b*r	b*	r	b*r
Planning (A)	.3124	.55	.17184	.1983	.66	.13086
Investigating (B)	.2318	.41	.09502	.0637	.60	.03822
Coordinating (C)	.1662	.39	.06843	.2780	.67	.18623
Evaluating (D)	.0512	.33	.01689	.0403	.55	.02215
Supervising (E)	.2610	.44	.11483	.1925	.72	.13861
Staffing (F)	-.1324	.36	-.04767	.2975	.67	.19932
Negotiating (G)	.2031	.40	.08125	.0103	.52	.00535
Representing (H)	.2275	.41	.09326	.2034	.45	.09155
		$R^2 = \Sigma b^*r =$	.59025		$R^2 = \Sigma b^*r =$	.81231
		R =	.76827		R =	.90128

Note: In vector notation, the R^2 formula used here is $r^2_{mult.} = \underline{b}^*\underline{r}$ (Winer 1971:127) which is equivalent.

These results show that there is a substantial difference in the percentage of the variance accounted for by these performance dimensions between the self ratings ($R^2 = .59$) and the superior ratings ($R^2 = .81$). This difference is significant at p .01 (F= 13.603).

The rating strategies of the self- and superior raters can be seen by examining the column titled b*. Superior raters clearly place considerable more weight on Staffing (F) behaviors than do the manager ratees themselves. In fact, in the self ratings, this dimension carries a negative regression weight. The difference in weights for this variable suggests that this performance dimension is less relevant or is less valued for the self raters than it is to the supervisors.

A similar difference is seen with the Coordinating (C) dimension where superiors have a regression weight of .2780 and self raters have .1662. Again, apparently this type of activity or behavior is either less relevant or less valued by the manager ratees themselves.

The self raters apparently value Planning (A) and Supervising (E) as being the most relevant dimensions of overall performance. These two variables have the largest beta weights for the self raters and have the most substantial contributions to the self R^2.

The contributions of each of the dimensions to the R^2 can be studied by observing the values in the column titled b*r. These values are the product of the standardized regression weights and the validity coefficients. As noted earlier, these values sum to the R^2 and, in the present situation where we are comparing ratings on the same individuals by two different sets of raters, these values can be considered to reflect the relative contribution of each dimension to the respective R^2 values.

The relative contribution of some of the variables is fairly comparable between the two rating policies; however, there are substantial differences in the relative contribution for Staffing (F) and Coordinating (C) between the two rating policies. These variables account for considerably more of the variance in supervisory ratings than in self reports. For the self ratings, Planning (A) and Supervising (E) make higher relative contributions to the overall prediction than they do in the supervisor's ratings. In this case, comparisons based on b*r values are the same as when the comparison was only with the b* weights.

The relative contribution values (b*r) must be interpreted with some caution. Ward (1962) and Darlington (1968) have objected to the use of this kind of measure when it is interpreted as an index of the "importance" of a variable in any test selection procedure or in rank ordering the "value" of variables. The numerical value of b*r will change as variables are added or dropped from the regression equation and certainly these values would not be stable across different samples. However, since the present use of these values is essentially descriptive (that is, to demonstrate that some difference in rating strategy exists between two classes of raters in a given set of empirical data) rather than inferential, its use can be both meaningful and potentially of considerable value. We may distinguish its use for this type of purpose by calling this product (b*r) the "relative contribution" to R^2, as opposed to any consideration of it as an index of the "independent contribution" of variables as used by Hoffman (1960, 1962).

The demonstrated difference in rating strategy may help us to understand the relatively low, although statistically significant, convergent validity of overall ratings reported by Heneman. It would appear that self and superior raters do not use the same rating strategies and that they place quite different values on the various dimensions of performance.

This may be a function of differences in experience or it might be a result of different organizational perspectives, as has been suggested by the work of Naylor and Wherry (1964) and Zedeck, et al (1974).

The substantial difference in overall prediction between the self and superior raters indicates that a considerably larger percentage of the variance in performance rating on the part of supervisors was accounted for in their assessment of the eight functional dimensions. Naylor and Wherry have observed that the R^2 values should be used in such situations "as a measure of the extent to which the policy has been captured (Naylor and Wherry 1964: 1)." Thus, we have not been able to capture as much of the rating policy of the self-raters as for the superiors. This seems to imply that other, unmeasured factors were entering into the self judgements of performance by the subjects in Heneman's study. Before speculating as to the possible relevant additional variables which might account for this difference, we need first to determine if a similar difference in policy is evident in other studies.

The same policy capturing procedure was applied to the data published by Lawler (1967) in the basic article outlining the multitrait-multirater approach. Lawler's Table 2 gives intercorrelations between Performance (A), Ability (B), and Effort-put-forth (C) in ratings by self, supervisor, and peers. Policy capturing of the rating strategies implicit in Lawler's multitrait-multirater matrix (using B and C to predict A) resulted in the data shown in Table 2.

Table 2

Prediction of Performance from Ability & Effort

Variable	Self			Peer			Superior		
	b*	r	b*r	b*	r	b*r	b*	r	b*r
Ability (B)	.38	.43	.1634	.40	.56	.2240	.33	.53	.1750
Effort (C)	.35	.40	.1400	.40	.56	.2240	.45	.56	.2520
		R^2 =	.30		R^2 =	.45		R^2 =	.43
		R =	.55		R =	.67		R =	.64

The most interesting information to emerge from this analysis is the completely balanced weighting of variables in the peer rating, which also yields the highest prediction (R = .67). Effort and Ability account for slightly more of the variance in peer ratings than in the supervisor's rankings, and substantially more than in the self-ratings. The marked reduction in R^2 for the self-ratings may be in part an artifact resulting from method variance; peer and supervisors accomplished rankings where the self reports were ratings (Lawler 1967: 374). However, Lawler apparently did not consider this difference in the way the data was gathered to be a serious problem.

The relative weights of ability and effort for the various types of rater's policies show some differences; self-raters weight ability slightly higher than effort, peers weight ability and effort equally, and the supervisors give considerably more weight to effort. This is a very interesting finding and one which has implications for both the scientific and practical use of the various types of ratings as performance criteria.

While peer rating policy is quite similar to the supervisor's in terms of the percentage of variance in performance rating accounted for, it is more similar to the self-ratings in terms of the roughly equal weighting of the two variables. Thus the rating strategy of peers is more like that of the self-raters even though peers have a much higher overall prediction. In terms of R^2, it is fairly obvious that there are other, unmeasured variables which are entering into or moderating the rating judgements; however, as seen previously in the reanalysis of Heneman's data, this is more so in the case of self-ratings than it is for superior or peer raters.

Discussion

The patterns which emerge from this restudy of Lawler's data are somewhat similar to the conclusions of Klimoski and London (1974) who felt that there were differences in how dimensions are evaluated by different types of raters. However, the explanation advanced by Klimoski and London was that the supervisors could not distinguish between effort and ability. The present data suggests that supervisors may be able to make that distinction but that they value effort somewhat more, in making their overall evaluations, than they do ability. Peers, on the other hand weight the factors equally.

The pattern of self-reports providing less prediction of the overall performance rating, which was evident in the reanalysis of Heneman's data, is repeated with Lawler's information. Thirteen to fifteen percent more of the variance in the overall rating can be predicted from the dimension ratings by superiors and peers than from the self-ratings. This replication of pattern across these two sets of data suggests that other, unmeasured variables are more frequently used by the individual when he judges himself.

It may well be, as Korman (1971) and Ghiselli (1967) have suggested, that the self concept of the individual is functioning as a moderator or as an additional factor when the individual judges himself. Nebeker and Mitchell (1974) felt that the individual's cognitive expectations might be the operant additional variable. It could be a number of things and this presents an interesting question for future performance appraisal research. Peers and supervisors appear to be able to make their judgement more on the basis of the taxonomy of variables or dimensions specified by the researcher but these two types of raters do not necessarily use the same rating strategy in weighting these variables into a composite overall performance rating.

These findings, if they can be further replicated, have some implications for the industrial and scientific use of self, peer, and supervisory ratings. For industrial use, if there is a substantial difference between the way a person judges himself on job performance and how his supervisor and peers judge him, then the use of self-reports might help to identify these differences and to serve as a starting point for the supervisor to help the individual view his performance realistically. Haire (1964) and Bassett and Meyer (1968) have observed that self reports can help get the individual involved and committed in the evaluation process. Porter and Lawler (1968) felt that the employee must take an active role in goal setting. Self appraisals may be a useful way to encourage this.

Haire (1964) has noted that the evaluation process is frequently something which "happens to" the employee rather than being a problem-solving, goal-setting tool. The present finding, which suggests that there are basic differences in evaluation strategies between self-raters and superiors, may explain why this is so; if superiors do not understand that a basic difference in perception of performance exists, they cannot help the employee to better understand his performance rating. Comparison of the supervisory appraisal with the individual's self-rating could make such differences very visible and lead to better understanding between supervisors and ratees.

For applied research purposes, the existence of such differences in rating strategy between self, peer, and supervisor ratings suggests that additional research is needed to identify the systematic differences in the various types of ratings. The data in the present study concerning peer ratings tends to further support recommendations made over a wide span of years that peer ratings may be a reasonable and useful performance criterion (Wherry and Fryer 1949; Hollander 1954; Shenk, Watson and Hazel 1973). The reanalysis of Lawler's data suggests that peer ratings represent greater prediction of overall rating, but that peers are actually using a rating strategy much more similar to self-raters than to the strategy of superiors. Peer raters share the same organizational perspective as the rated employee where supervisors are at a different level in the organization and may well have quite different values. Further peer ratings seem to be more objective than self ratings, at least in terms of making greater use of the rating dimensions or variables specified by the researcher. Further research is needed to increase our understanding of the potential of peer ratings as a performance criterion.

Conclusions

This investigation was undertaken in an attempt to understand the continuing difficulty researchers have encountered in demonstrating discriminant validity of constructs via the multitrait-multirater approach. It became evident upon restudy of the reports of Heneman (1974) and Lawler (1967) that one probable cause of this difficulty lay in the inclusion of an overall performance rating in the multitrait-multirater matrix, which results in an inappropriate comparison of convergent validities with values which are themselves validity coefficients in their own right, rather than being the chance correlations between variables (heterotrait correlations) expected in the method of Campbell and Fiske (1959). Elimination of such global performance measures from the comparisons made to demonstrate discriminant validity is recommended. However, it was demonstrated that such overall performance measures may be of value as criteria in a policy capturing model; comparison of the different rating strategies of classes of raters can be a valuable additional source of information from the multitrait-multirater data.

Such a procedure also permits the comparison of R^2 values for the various classes of raters. If substantial differences in overall prediction do exist, as was the case in this study, then the normal comparisons made within the multitrait-multirater matrix must be interpreted with some degree of caution, since other unknown variables are operative in one set of ratings but not in the other. The examination of R^2 values derived from the heterotrait-monorater triangles of the matrix can be of value in its own right, since it may lead to quite different conclusions than would be made using only the multitrait-multirater approach.

References

Arvey, R. D. & Hoyle, J. C., A Gutman approach to the development of behaviorally based rating scales for systems analysts and programmer/analysts. Journal of Applied Psychology, 1974, 59, 61-68.

Bassett, G. A., & Meyer, H. H., Performance appraisal based on self-review. Personnel Psychology, 168, 21, 421-430.

Borman, W. C. The rating of individuals in organizations; an alternative approach. Organizational Behavior and Human Performance, 1974, 12, 105-124.

Campbell, D. T., & Fiske, D. W., Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 159, 56, 81-105.

Darlington, R. B., Multiple regression in psychological research and practice, Psychological Bulletin, 1968, 69, 161 - 182.

Ghiselli, E. E., Interaction of traits and motivational factors in the determination of the success of managers. Journal of Applied Psychology, 1968, 52, 480 - 483.

Haire, M. Psychology in Management. New York: McGraw-Hill, 1964.

Heneman, H. G. III, Comparisons of self- and superior ratings of managerial performance. Journal of Applied Psychology, 1974, 59, 638 - 642.

Hoffman, P. J., The paramorphic representation of clinical judgement. Psychological Bulletin, 1962, 57, 116-131.

Hoffman, P. J., Assessment of the independent contributions of predictors. Psychological Bulletin, 1962, 59, 77 - 80.

Hollander, E. P. Buddy ratings; military research and industrial implications. Personnel Psychology, 1954, 7, 385 - 393.

Kavanagh, M. J., MacKinney, A. C., and Wolins, L., Issues in managerial performance: multitrait-multimethod analysis of ratings. Psychological Bulletin, 1971, 75, 34 - 49.

Klimoski, R. J., and London, M., Role of the rater in performance appraisal. Journal of Applied Psychology, 1974, 59, 445 - 451.

Korman, A. K. Industrial and Organizational Psychology. Englewood Cliffs, New Jersey: Prentice-Hall, 1971.

Lawler, E. E. III, The multitrait-multirater approach to measuring managerial job performance. Journal of Applied Psychology, 1967, 51, 369 - 381.

Mahoney, T. A., Jerdee, T. H., and Carrol, S. J., Development of Managerial Performance: A Research Approach. Cincinnati: South-Western, 1963.

Mahoney, T. A., Jerdee, T. H., and Carrol, S. J., The job (s) of management. Industrial Relations, 1964, 3, 97 - 110.

Naylor, J. C. and Wherry, R. J. Sr., Feasibility of Distinguishing Supervisors' Policies in Evaluation of Subordinates by Using Ratings of Simulated Job Incumbents. Technical Report PRL-TR-64-25, Personnel Research Laboratory, Aerospace Medical Division, Lackland Air Force Base, Texas, October 1964.

Nebeker, D. M., and Mitchell, T. R., Leader behavior: an expectancy theory approach. Organizational Behavior and Human Performance, 1974, 11, 355-367.

Porter, L. W. and Lawler, E. E. III., Managerial Attitudes and Performance. Homewood, Ill.: Irwin, Inc., 1968.

Shenk, F., Watson, T. W. (Sgt., USAF), and Hazel, J. T., Relationship Between Personality Traits and Officer Performance and Retention Criteria. Technical Report, AFHRL-TR-73-4. Personnel Research Division, Air Force Human Resources Laboratory, Lackland Air Force Base, Texas, May, 1973.

Ward, J. H. Jr., Comments on "The paramorphic representation of clinical judgements". Psychological Bulletin, 1962, 59, 74 - 76.

Wherry, R. J. (Sr.) and Fryer, D. H., Buddy Ratings; popularity contest or leadership criteria? Personnel Psychology, 1949, 2, 147-159.

Williams, W. E., and Seiler, D. A., Relationship between measures of effort and performance. Journal of Applied Psychology, 1973, 57, 49 - 54.

Winer, G. J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1971 (2nd Edition).

Zedeck, S., Imparato, N., Krausz, M., and Oleno, T., Development of behaviorally anchored rating scales as a function of organizational levels. Journal of Applied Psychology, 1974, 59, 249 - 252.

A Survey of Current Practices In Occupational Certification and Licensing Testing

Michael D. Hiscox and Dean H. Nafziger
Northwest Regional Educational Laboratory

BACKGROUND INFORMATION

A study program and test collection project was recently completed by the Clearinghouse for Applied Performance Testing at the Northwest Regional Educational Laboratory (NWREL). This work was sponsored by the United States Office of Education at the request of the Defense Activity for Non-Traditional Educational Support (DANTES).

The project focused on two separate, but related, objectives: compilation of self-study programs leading directly to certification or licensing and development of a survey of occupations to provide both general and specific information on licensing and certification procedures for those occupations in which vocational competence is determined through examination.

PROCEDURES

Information Solicitation and Collection

To determine the availability of certification tests, NWREL staff contacted over 1,000 potential information sources by telephone, first-class mail or personal interview. University and municipal libraries provided available written references, while three computer searches supplied abstracts of significant educational and vocational articles.

The telephone information search yielded mixed results. In some cases, project personnel were already aware of the existence of an appropriate course or examination procedure, and the telephone call was merely a request for more detailed information. Phone calls in such cases were quite effective. But when a call was made to determine the existence of an appropriate program, it often proved less helpful.

Calls to state and municipal government agencies were particularly ineffective, possibly because of the large size of those agencies and the extensive division of responsibility among agency staff. On the other hand, most professional organizations and Federal agencies had specific, well-established programs and were therefore capable of providing adequate, complete information with minimal delay and confusion.

At the start of the project, reference materials were needed to provide information for the occupational survey entries. In addition, it was necessary to determine what previous work had been done in relevant fields. Two federal publications of obvious value--The Dictionary of Occupational Titles[1] and the Military Civilan Job Comparability Manual[2]--provided information useful in preparing the compilations. However, The Occupational Outlook Handbook[3] was the project's most useful reference. Its entries not only contained descriptions of the nature of the work and training requirements for each occupation, but also listed sources of additional information. Project staff used these listings to check the completeness of information they collected. The Handbook often provided general information which was not immediately available from other sources.

A few additional references were used. The Dictionary of Professional and Occupational Licensing in the United States[4] and Equivalency and Proficiency Testing[5] were useful references against which the completeness of the DANTES occupational survey could be checked. Other references provided addresses for organizations, state agencies and unions.[6,7,8,9] Occupational Licensing: Principles and Practices[10] offered an indepth study of current licensing concepts and processes. No other references were found which provided information directly related to project development goals.

Computer literature searches were conducted by NWREL Information Center personnel using the ERIC, AIM (Abstracts of Instruction Materials) and ARM (Abstracts of Research Materials) systems. Several hundred references and their abstracts were obtained through these searches, but only a small percentage of the titles were appropriate, and the large majority of those provided only minimal information. Most references related to apprenticeship or manpower programs, or to individualized vocational programs in the educational setting. In summary, the computer searches were useful only in providing background information.

The project's mailing program was the most valuable method of collecting information. A form letter was sent to a wide variety of agencies, organizations, and individuals who might have information to contribute. Information on both independent study courses leading to licensure or certification and the licensure or certification procedures themselves was requested.

Project staff took special care to develop a mailing list which thoroughly covered the broad spectrum of potential information sources. The original mailing list selection procedures are outlined below.

1. State Governmental Agencies. For each state, approximately five agencies who might have relevant information were selected from the 1974-75 National Directory of State Agencies.[7] Selections typically included Departments of Education, Commerce, Labor, Consumer Affairs, and so forth. Slightly over 300 contacts were made through this procedure.

2. Other Governmental Agencies. Several municipal and federal agencies were contacted. In a few cases the intent was to gather general information, but more often it was to obtain information about specific known procedures. It was for this latter purpose that the licensing divisions of the Federal Aviation Administration, Federal Communications Commission, and the United States Coast Guard were contacted. Fewer than ten agencies were contacted within this group.

3. Home Study Schools. All member schools of the National Home Study Council were contacted. (Letters were not sent to schools listed as subsidiaries or divisions, but in all cases the parent organization was contacted.) Other correspondence schools listed in Lovejoy's Career and Vocational School Guide[8] were also contacted. In all, 100 schools received the information request.

4. Colleges and Universities. All of the approximately 75 member schools of the National University Extension Association (NUEA) received the information solicitation.

5. Companies. A request for information was sent to each of the 100 largest United States industrial corporations specified in the appropriate 1975 Fortune Magazine Double 500 listing.

6. Labor Organizations. Through examination of the 1974 National Trade and Professional Association's Yearbook[6] the names of approximately 70 labor unions were obtained; organizations which were, from name, presumed to be labor unions were selected from a listing of groups reporting annual budgets of over $1,000,000.

7. Professional and Trade Organizations. This category originally contained 70 organizations which were selected because they seemed potential sources of relevant information. The selection was made on the basis of information from the Occupational Outlook Handbook[3] and Lovejoy's Career and Vocational School Guide[8] as well as prior knowledge of the project staff.

8. Other Groups and Individuals. The solicitation letter was also mailed to a small number of potential contributors who did not fall into one of the groups listed above. Names of such organizations and individuals were often suggested during the course of phone calls or were listed in relevant references and materials. These contacts--less than 15 in all--included professional educators knowledgeable in the vocational licensing field and a small number of centers for vocational education or research.

The form letter used to solicit information asked that the recipient provide project staff with names of additional contacts who might provide relevant information. Many did so, providing NWREL with over 100 additional information sources. It was the policy of project staff to follow up on all suggested contacts--occasionally by phone, but more frequently through use of the same solicitation letter. State agencies most often provided these followup contacts; when the original agency was an inappropriate source of information, their response usually contained the suggestion of a more probable source. Frequently the state agency itself forwarded NWREL's letter to a different government department.

A few individuals and universities and several professional organizations provided suggestions for additional information sources. Responses from state governmental agencies and from professional and trade organizations were generally supportive of the project objectives; therefore, such organizations often attempted to provide sources of information when they themselves were unable to help.

Table 1 summarizes the scope and success of the information solicitation phase in each of the classifications described above.

In general, information provided by state governments was wide ranging, but seldom detailed enough to be of significant value. There were exceptions, however, as some states had previously compiled single sources of complete information on licensed occupations within the state.

Most information provided by colleges and universities took the form of printed extension service catalogs and bulletins; this information was detailed but not particularly applicable. Their response rate was moderate. Few companies or labor organizations responded and those which did provided no important information. Response from home study schools was also very low.

The few non-state government agencies contacted responded well to specific information requests. Professional and trade organizations were usually able to either provide useful information and appropriate information sources, or state that the information requested did not exist. Their response rate was fairly high, and the detailed booklets, pamphlets and brochures provided by those organizations were instrumental in providing both occupational and examination information.

Overall, the information solicitation phase was judged successful. Adequate information was available on all occupations and programs NWREL wished to include in the compilation, and the information received appears to cover the majority of existing occupational examination processes.

Table 1

MAILING LIST CONSTRUCTION AND RESPONSE

Classification	Original* Contacts	Subsequent* Contacts	Total* Contacts	Total*# Responses
State Governmental Agencies	300	75	375	175
Other Governmental Agencies	5	0	5	5
Home Study Schools	100	0	100	5
Colleges and Universities	75	5	80	25
Companies	100	0	100	5
Labor Organizations	70	0	70	5
Professional and Trade Organizations	70	30	100	50
Other Groups and Individuals	10	5	15	5
TOTAL	730	115	845	275

* All figures are approximated slightly.

# This category tabulates all written responses received, including materials, information, suggestions for additional contacts, and statements that information was not available.

PROJECT FINDINGS

Evaluation of the Information Collected

The results of the information search permit some observations concerning licensing and certification procedures and related independent study courses. These observations are supported by the thorough nature of the project's information collection process; however, until verified by widespread empirical research, they must be regarded as tentative.

Numerous independent study courses, covering a wide range of subject areas, are available. Proprietary schools, colleges, universities, junior and community colleges, and various professional and trade organizations all offer self-study, correspondence coursework. Courses range from elementary through post-graduate levels, and cover everything from highly practical skills like welding to theoretical science and math. Project staff readily identified several thousand such courses.

Yet very few of these courses lead to recognition of achievement through award of certificate, license, or academic degree. Most individual courses are apparently intended to increase knowledge or skill in one restricted area; seldom are they coordinated to form a total "program." Project staff therefore realized that very few independent study courses were suitable for inclusion in a catalog of self-study programs--particularly a catalog of programs providing preparation for certification or licensing examinations. A few programs offer "preparation courses" for real estate, accounting, engineering, or other professional examinations. In addition, the American Watchmaker's Institute has a self-study course and related certification test in watchmaking and a few proprietary schools have curricula which lead to various licenses--notably Federal Communications Commission permits. Yet it would probably be correct to state, if our sampling is representative, that not one of a hundred independent study courses is directly related to a licensing or certification examination.

Colleges and universities provided a large part of the independent study information. Even so, their response rate was not as great as anticipated. Even more surprising was the very small response received from home study schools. Of more than a hundred schools contacted only five responded. Clearly home study schools have materials available for dissemination. Therefore, project staff concluded that either the schools offered no courses appropriate for inclusion in the survey, or perhaps the compilation procedures, the sponsoring body, or some other aspect of the project made program evaluation unattractive to the majority of these organizations. The DANTES decision not to include proprietary schools in the compilation eliminated the problem of obtaining information from these sources.

Project staff were considerably more successful in compiling examination information. Although state governmental agencies were seldom able to provide information about tests with nationwide applicability, these agencies were often able to supply fairly detailed information on requirements for a few occupations within a particular state (e.g., plumbers in Washington, elevator inspectors in Maine). Though they were usually unable to generalize licensing requirements to other states, that is not to say that the contacts were not useful. The large volume of responses often made it possible for project staff to make required generalizations themselves. As previously mentioned, among the most helpful contributions were the compilations received from about ten states. These documents were very valuable in developing survey entries on the licensed occupations.

Companies and labor unions provided almost no input to the compilation; the few that did respond offered no relevant information. However, various supplemental readings and interviews suggested that most company programs were actually not appropriate for the survey in any event, since they were not available nationwide, nor were they available to people outside the employ of a specific company. Labor unions could have provided much information regarding how union membership and journeyman status constitute a type of certification--a concept which affects several million workers in union-oriented trades. However, such information was outside the focus of the compilation. The response rate from companies and unions was so low that it is logical to conclude that many requests did not reach appropriate sources in the first place (because of inadequate addresses), or that none of these sources had any pertinent information.

Most useful information came from the various professional organizations, who provided all certification and a substantial amount of licensing information used in the compilation.

Comments on Licensing and Certification

Many occupations are in some way regulated by state, federal, or local governmental agencies; such regulation is generally termed licensing. Licensing requirements most often originate through specific legislative action. When a professional organization, rather than the government, sponsors the regulatory process, it is called certification. Certification procedures are most often used to determine the occupational competence of workers, and to identify those who do not meet minimum standars. Completion of the certification process advertises a worker's abilities to employer, colleagues, and customers. (A few occupations offer credentialing through the use of a national registry; qualified personnel receive recognition through listing in a directory available to co-workers and employers. This study found no difference between the intent or effects of the registration procedure and that of certification.)

Certification provides some assurance that a worker has met certain minimum standards in the field. This is not to say that only certified personnel are competent. There are a wide variety of reasons (including cost and opposition to the certification concept) why qualified personnel may choose not to be certified. Nor is the value of certification consistent; certification plays a very important role in some fields, while in other fields, it has almost no impact on obtaining, maintaining, or improving employment opportunities.

It should be remembered that, regardless of its importance, certification is ultimately a voluntary process, while licensing procedures, no matter how unrelated to occupational proficiency, are (at least in concept, if not in enforcement) mandatory. A vast number of licensing procedures have developed over the years, some to protect the public welfare, but many as revenue producing or registration procedures, or means of restricting entry into a particular occupation to the benefit of current members. During the survey compilation project staff found such wide variation in licensing requirements for the same occupation from state to state that "necessary occupational competency" (as ascertained by licensure) cannot possibly be equivalent throughout a licensed trade or profession. The study Occupational Licensing: Principles and Practices[10] deals in depth with questions related to the value of, need for, and potential for improvement in the current maze of governmental licensing procedures.

Certification is always nationally uniform in character, while state licensing (especially outside highly professional occupations--such as law or medicine) is generally unstandardized and inconsistent. While licensing is a mandatory process, it was not within the scope of the project to determine how closely licensing requirements relate to job tasks; the disparity among states' requirements suggests that there is reason to question that relationship. Although certification procedures seem more logical (due in part to their uniformity) there is some question regarding the value and necessity of the process. Once again, that question lies outside the project's focus.

TENTATIVE CONCLUSIONS DERIVED FROM THE PROJECT WORK

Project staff have reviewed a great amount of information in the preparation of the DANTES documents; that review allows the formation of several tentative conclusions related to occupational certification and licensing. Various references and interviews with a number of individuals and organizations provided ancillary information which helped in formulating and substantiating some of these conclusions. It would appear that a wide variety of practices and concepts in occupational credentialling could benefit from additional evaluation.

1. With only a few exceptions, independent study courses do not seem to allow completion of a comprehensive program of study without a significant amount of additional non-independent training, either in the form of supervised job experience or institutionally-based educational coursework. It is generally impossible to obtain licensure or certification solely, or even primarily, through the completion of self-study courses. Most independent study programs, even those which result in degrees or educational certificates, do not relate directly to any type of licensing or occupational certification program. Any attempts to compile listings of these "directly-related" courses will be limited, as was the NWREL effort, by the apparent small number of existing programs. With the exception of several proprietary school courses, project staff are aware of only one totally appropriate program, the University of North Carolina's correspondence course in dental assisting. This non-credit course is approved preparation for the American Dental Assistant's Association certification test. A few other programs (less than five) are somewhat appropriate for consideration.

2. Considering the vast number of occupations, only a small number have any type of competency-based examination procedure which clearly evaluates a person's ability to satisfy occupational responsibilities. Over 2,000 occupations have licensing or certification procedures of some sort, but only a fraction of these involve demonstration of the applicant's competency through examination.

3. In some occupations (notably the building and industrial trades) and on some job sites, labor unions have, in a variety of ways, greatly influenced job entry requirements. This employment-limiting process is similar in effect to licensing or certification procedures. Various readings and interviews suggest that unions are typically in favor of maintaining current processes rather than encouraging increased criteria-related competency testing as a measure of a worker's skills.

4. The information collected indicates that most states sanction educational institutions as "licensing agencies;" graduation from a training institution accredited in a given field often satisfies a large portion of the licensing requirements. (As an obvious example, no amount of experience or successful examination performance will allow one to become a physician without graduation from an approved medical school. This procedure affects many occupations, from law and dentistry to barbering and dry cleaning.)

5. Ability to gain employment or advancement in a field can rarely be predicted solely on the basis of satisfying certification or licensing qualifications. Credentialling usually varies in importance, depending on an individual employer, geographic location, job market, and so on.

6. Only a very limited amount of research seems to have been done in the area of occupational licensing and certification. Of particular concern is the complex issue of how closely certification or licensing relates to actual job competency and what benefits are derived from the existence of these processes.

7. Many governmental agencies and some professional organizations welcome attempts at compiling and examining certification and licensing procedures. Most communications to the project staff were supportive of the DANTES undertaking.

8. Many states have not centralized or even identified occupational licensing information in a manner which would make it relatively easy to obtain answers to general or specific licensing responsibilities; it is far more common to find anywhere from ten to fifty separate departments and boards each maintaining responsibility for one, or several, related occupations. The project staff often had difficulty locating an appropriate information source within the state government; it is likely that this situation would create difficulty or inconvenience for those seeking assistance with routine licensing problems.

9. The certification procedure often seems to be primarily an attempt by current members of an occupation to increase the professionalism and status of their field. Certification procedures are, in almost all cases, initiated from within an occupation rather than by any outside legislation or group action. Certifying agencies occasionally mentioned the need to bring increased recognition to successful practitioners of an occupation and/or to maintain high standards of competency within the field.

10. Certification is most predominant in the allied health occupations. It is hypothesized that this is partly due to the importance and highly technical nature of the work performed (which makes both "recognition" and "maintenance of standards" viable rationales) combined with the obvious need to protect the public welfare. Certification is a significant part of many health fields; salary, advancement potential, and even ability to obtain work in the first place are, because of employer acceptance of the certification concept, often dependent on successful completion of the credentialling process.

11. State licensing requirements vary so greatly that it is easy to question the validity of the criteria on which they are based. For example, 1,000 hours of classroom instruction are required for a cosmetology license in Texas, while 1,800 hours are required in Arizona. Does this imply that the cosmetologists in Arizona are considerably better qualified, or that the training in Texas is significantly more effective? More likely it suggests that the requirements themselves are somewhat arbitrary, that they are not the result of a systematic pairing of the skills with the curriculum. Information collected for the survey revealed a multitude of such discrepancies; in only a few licensed occupations was there substantial uniformity among state licensing requirements.

12. Information collected suggests that licensing and certification tests may not generally be constructed under the same rigid criteria which would be used for educational tests of equivalent importance. In all the test descriptions examined, project staff found no mention of validity or reliability checks. Test procedures were often not standardized. Perhaps most significant was the fact that testing in many occupations was of the norm-referenced type in which a candidate must obtain a percentile score which exceeds a specified "passing" score before being granted the credentials. These cutoffs frequently seemed unrelated to any competency-based criteria and appeared extremely arbitrary. Several occupational examinations have been developed under rigorous criteria that help ensure their appropriateness and value. Yet many other tests do not necessarily succeed in determining the true competence of the candidate. They are seldom rigorously constructed tests with a strong correlation to a proven set of essential abilities. Many appear to have been established solely as a result of legislative pressures, or the opinions of members of the occupation regarding how much a person should know about what topics. Such persons hold informed opinions, to be sure, but tests constructed without empirical justification for their content may be considered a significant weakness in the overall credentialling process.

This is in no way meant to imply that these tests are inappropriate--their appropriateness is a question which requires extensive research before even tentative conclusions can be drawn. Nor is it meant to suggest that the test developers were not qualified. The implication is simply that few licensing or certification agencies can empirically demonstrate that their credentialling tests do indeed differentiate between individuals on the basis of occupational competency.

In conclusion, this project has gathered a wide variety of information in the field of occupational licensing and certification. Yet because this area has not been extensively investigated, additional research would seem to be in order.

REFERENCES

[1] United States Department of Labor, Bureau of Employment Security. Dictionary of occupational titles. Washington: U.S. Government Printing Office, 1975, 2 vols. plus supplement.

[2] United States Department of Defense, Office of the Assistant Secretary of Defense. Military-civilian job comparability manual. Washington: U.S. Government Printing Office, undated.

[3] United States Department of Labor, Bureau of Labor Statistics. Occupational outlook handbook. Washington: U.S. Government Printing Office, 1974.

[4] Angel, J. L. (Ed.) Dictionary of professional and occupational licensing in the United States. New York: World Trade Academy Press, 1970.

[5] National Institute of Health, Division of Allied Health Manpower. Equivalency and proficiency testing: A descriptive compilation of existing testing programs in allied health and other health occupations, with an annotated bibliography. Bethesda, Maryland: Author, 1971.

[6] Colgate, C. (Ed.) National trade and professional associations of the United States and labor unions. Washington: Columbia Books, 1974.

[7] Vellucci, M. J., Wright, N. D., & Allen, G. P. The national directory of state agencies, 1974-75. Washington: Information Resource Press, 1974.

[8] Lovejoy, C. E. Lovejoy's career and vocational school guide. New York: Simon and Schuster, 1973.

[9] New York State Bureau of Occupational Education Research. Occupations needing credentials. Albany: The University of the State of New York, The State Education Department, December, 1974.

[10] Shimburg, B., Esser, B. F., & Kruger, D. H. Occupational licensing: Practices and policies. Washington: Public Affairs Press, 1973.

PROGRAM LEVEL CREDIT AWARDS FOR OCCUPATIONAL COMPETENCIES: MILITARY-CIVILIAN LINKAGES

Duane O. Geiken, PhD
DANTES
Pensacola, Florida

Most people in government service are aware of the continuing pressure from Congress and government officials for meaningful interface between civilian and military education and training. There are many reasons for this, but the main reason is the belief that if the military relies more heavily on existing civilian education and training agencies, there is going to be greater cost effectiveness. Defense Activity for Non-Traditional Education Support (DANTES) is an important part of this movement as it relates to voluntary education opportunities for military personnel. In general, however, the climate for meaningful interface with civilian education and training has improved considerably in recent years to a point that the more traditional civilian programs do in fact interface very well with much of the formal education and training which takes place under the Military.

Over the past decade, the most dramatic development in civilian education and training has been development of community college systems across the country and the emergence of community colleges as a major force in higher education in the United States. The importance of community colleges in the present, as well as the future, has not yet been fully recognized by the Military. Most community colleges have placed emphasis on career education and job entry training programs and with this emphasis has come interest in performance-based curriculum in occupational training programs. This has been particularly true for those colleges which have defined their primary mission as one to meet needs of business and industry in their communities and have moved quickly to establish direct relationships with employers. Generally, this practice has proved beneficial to military personnel, especially those who have been stationed near community colleges. State governments also have been instrumental in promoting greater involvement with employers, including the Military. Most state departments of education have mandated that advisory committees or councils composed of persons from industry be established by colleges and technical institutes to insure employer participation in designing curriculum and training which will meet performance expectations of business and industry. The federal government, through the U. S. Office of Education and the Department of Labor, also has required institutions operating government funded training programs (MDTA and the successor CETA training programs) to establish advisory committees composed of employers from within the community. The federal government, even more than local college governing boards and state coordinating agencies, has insisted that curriculum and training be designed at levels of competency based on measurable performance objectives acceptable to employers.

These developments in civilian education and training over the past decade have been, and will continue to be, highly significant to military education and training because performance-based technical training and evaluation of individual competencies are the foundation of most military training and career advancement programs. In my opinion, there is now general acceptance by academic institutions of most of military education and training due in large measure to dramatic changes which have occurred in post secondary education through community colleges and their emphasis on career education. The old claims that civilian training and education cannot be as clearly defined as the military in terms of performance criteria, and that civilians' lack of resources and line-type organization necessary for successful implementation of performance-based training, now come from only isolated pockets of resistance. What is most evident is a climate of cooperation which has fostered unprecedented linkages between military and civilian education and training. We see today that almost every military installation which is not at a remote location has established linkages with local education systems, community colleges, and in some cases with nearby universities. It is not uncommon to see as many as ten institutions offering college courses on a military base. Overseas and at remote locations, voluntary education services have been made available through contract arrangements with civilian institutions. There have been other major developments. The services have taken the initiative in establishing programs, institutions and other vehicles which promote military-civilian linkages, (i.e., Community College of the Air Force, Navy Campus for Achievement, Project AHEAD). Under joint sponsorship of national higher education associations and the military, the Servicemen's Opportunity College was established. Equally important are activities of the American Council on Education through its Office of Educational Credit in recommending levels of credit awards to civilian institutions for formal military education and training programs. Other projects, such as the recent agreement between the Federal Bureau of Apprentice Training and the Army to recognize military training for apprenticeship requirements, and the study of credit awards by MOS classifications, are highly significant in further expansion of civilian-military linkages in education and training.

It is natural that DANTES focus attention on competency testing and its potential for awarding credit-by-examination in occupational fields. Established a little over a year ago to continue the testing program formerly sponsored by the United States Armed Forces Institute (USAFI) and to make a transition from the in-house USAFI correspondence course program to reliance on independent study and extension courses offered by accredited colleges and universities, DANTES is in the forefront of the continuing effort to interface civilian and military education and training. One of the most visible deficiencies in the service's voluntary education program for military personnel is the lack of opportunities to obtain credit in associate degree and certificate occupational programs through non-traditional means. This was understandable for several reasons: voluntary education programs sponsored by the military have focused primarily on traditional education and training recognition in academic fields; there are practical administrative problems associated with demonstrating attained skills either through actual work on-the-job or through simulated circumstances; most attempts to gain acceptance for pencil and paper tests to measure competencies in occupational fields have not met with acceptance from either training institutions or employers; trade associations (professional and para-professional agencies, and others) interested in promoting credit-by-examination in occupational fields have generally avoided direct involvement because of vested interest accusations and related claims of self interest; education and training institutions have relied almost exclusively on challenge tests along traditional occupation course patterns included within established occupational programs and this practice has inherent problems not only because of the provincial nature of such challenge tests but also because these tests are difficult to articulate with competencies of individuals who come either from different training programs or who have attained their competencies primarily on-the-job with little or no formal training. All these problems and many other subliminal factors have hindered greatly progress in credit-by-examination in occupational fields.

At the national level, the most significant advance in occupational competency testing has taken place regarding teachers seeking state certification to teach in trade-industrial fields of occupational education. Several years ago, under a project grant from the U. S. Office of Education, the Graduate School of Education, Rutgers University, developed a consortium for Occupational Competency Testing of Trade and Industrial/Technical Teachers. The mission of this consortium was to reconcile certification problems for persons already employed as vocational instructors or those wishing to become employed but who lacked traditional credential requirements of work experience and/or training in their occupational fields. From this original project emerged the National Occupational Competency Testing Institute (NOCTI) which until recently was located within the campus complex of Educational Testing Service, Princeton, New Jersey. Today, NOCTI is a consortium of forty-four states represented by their respective education and training agencies, or by colleges and universities designed by the appropriate state official. The following features of NOCTI eliminate or minimize most of the problems which have been associated with competency testing:

1. The investment of hundreds of thousands of dollars by the federal government and subsequent financial support and staff involvement of almost

every state education agency has not only established a quasi public image for NOCTI but has also provided the degree of stability necessary for continued operations and further advances.

2. NOCTI Administrative Board establishes policies and procedures regarding test development and approval, operational services, and administration of the NOCTI testing program. Because this Board is comprised of representatives of public agencies and institutions, it cannot have vested interests but rather a strong public interest.

3. The twenty-four occupational fields for which NOCTI has developed examinations represent a broad spectrum of trade/industrial programs common to most post secondary institutions and fortunately also have corresponding career specialties in the Military.

4. Examinations are most attractive to employers for in addition to a two-hour written test, the principle feature of a NOCTI examination is the four- to six-hour "hands on" performance test which stresses both process and the quality of the product worked on by the candidate. The heavy emphasis on performance was the result of direct involvement of persons from business and industry in construction of examinations and insistence that tests measure what has been learned during candidate's working experience rather than measuring classroom or simulated shop learning.

5. More than twenty test centers have been established across the nation and an outstanding cadre of NOCTI test coordinators has been developed to conduct and supervise test administration twice annually under carefully designed and regulated conditions which guarantee the quality of NOCTI testing.

6. After two years of operations, more than sixty colleges and universities with recognized teacher training programs have extended credit or advanced standing to successful candidates of NOCTI examination program.

7. Several state departments of education have adopted NOCTI examinations as a requirement for teachers to prove occupational competency in trade/industrial fields.

These general characteristics and specific features of the NOCTI examination program have generated considerable interest among those who see other applications of the NOCTI program beyond its present use for purposes of teacher certification. The NOCTI Board and its Director have responded positively to the possibility of expanding the NOCTI program beyond its present limited mission. The most interest in utilization of present NOCTI examinations and possible adaptation for other purposes has come first from state governments who see application in evaluating the quality of persons being trained under programs funded by the state and federal monies, and second from community college systems in major metropolitan areas which have been leaders in extending credit-by-examination in academic fields through the College Level Examination Program (CLEP). It is community college interest that attracted DANTES to NOCTI.

After appropriate investigation and dialogue with community college personnel (including instructors, directors of occupational programs, college presidents, and system chancellors), there was concensus that the full potential of the NOCTI examination program should be explored. It was concluded by DANTES staff that a pilot project involving NOCTI was clearly within DANTES' mission to promote non-traditional education as part of voluntary education programs of military services. It was concluded also that a high-quality occupational competency testing program would be another important step toward greater interface between civilian and military training programs. The pilot project contract has been approved and a first meeting has been held to plan for implementation.

The potential of competency testing as another option for military personnel is far reaching. Formal meetings and informal discussion which have taken place over the past year have developed some interesting possibilities. The most attractive to many (and most threatening to others) is a departure from credit awards on a course basis to a more flexible practice of awards which abandons traditional course and shop identification and moves to a program award of credit. To illustrate how this might work in practice, let's relate it to a Mechanical Technology Associate in Applied Science degree program offered at the community college or technical institute level. The structured program normally requires at least 60 semester hours of formal instruction and laboratory with 15 to 24 of those credit hours required in general education and related courses. The basic core curriculum requires various courses in mathematics for machine technologists, manufacturing process, materials, drafting and design, metallurgy, machine design, machine operation, and other technical courses in the machine trades field. Using the present NOCTI examination (which is designed to evaluate knowledge and skill acquired on the job at the competency level of Journeyman), the institution would extend credit to a candidate up to 75% of the credits required for completion of a degree or certificate in that field. In this instance, maximum program credit award could go as high as 45 semester hours if the candidate demonstrated exceptional competency of his knowledge of mechanical technology on the written part of the NOCTI examination and high skills proficiency on the "hands on" performance test. Comprehensive analysis and review of individual examinations might result in lesser credit awards by the college or technical institute. Institutions which are willing to accept this flexible approach to granting credit will offer an additional option to its client population, especially the thousands of military personnel who have not yet received recognition or college credit for attained competencies in their specialized occupations.

The potential for military-civilian linkage under such a flexible credit award procedure, without regard to course structure, is great. There are approximately 2,500 education and training programs in the military which have a substantial degree of similarity to occupational programs offered by community colleges and technical institutes. Over 1,800 of these military programs are in engineering and industrial related military job categories (distributed percentage-wise between the services roughly 45% Air Force, 28% Army, 18% Navy, 6% Marines, and 3% other). These statistics reflect tremendous potential for the Military especially if recognition is extended to other skill levels lower than Journeyman level.

Established levels of proficiency (or skill) now used within the military do not deviate too far from levels of proficiency which have been established in many community colleges. In the case of the Air Force, 7 Skill Level (which would approximate Journeyman status) and the present NOCTI examination program appear to have a direct relationship. However, it may be feasible to relate 5 Skill Level and 3 Skill Level to certificate programs offered at community colleges. It is hoped that one or more levels of competency below the Journeyman level can be identified and that NOCTI tests either can be adjusted to measure such levels or that other criteria can be established to extend measurements of competency to lower skill levels.

The pilot project which will be conducted this year jointly by DANTES and NOCTI will involve all four military services and several community college systems and public universities. Some of the objectives of this pilot are:

1. To identify those characteristics of a test candidate now in the military which indicate reasonable degree of success in taking the present Journeyman level examination (i.e., training received, on-the-job training, work experience, level of skill rating, etc.).

2. For each occupational field, to identify "shortfalls" or deficiencies between military skill expectations and the civilian skill expectations reflected in NOCTI testing.

3. Explore the feasibility of conducting NOCTI examinations at military installations and the extent to which a network of military testing centers, similar to the existing network of civilian testing centers, would be possible.

4. Establish procedures for successful candidates to enroll in programs of institutions which will grant candidates credit-by-examination in occupational programs.

5. Develop information on the competencies of persons who complete occupation programs in community colleges and technical institutes at certificate levels and associate degree levels. While this will be primarily an activity of NOCTI and selected institutions and state departments of education, this information is basic to establishing other levels of competency testing within the NOCTI program below the level (Journeyman) of existing tests.

6. Make studies of unit costs and cost effectiveness of competency testing in relation to other programs available to military personnel through the voluntary education programs of the four Services.

There is a solid base of support within the Military for credit-by-examination in occupational fields and similar support from the more progressive institutions for increased activity and aggressive promotion. However, there is only guarded interest at best on the part of others who tend to cling to more traditional methodology. It should be recognized that there is outright resistance from some educators but more so from labor union people, mostly in the local organizations, who see competency testing as a threat to apprenticeship training programs. The key to the success of this NOCTI pilot project and other activities in competency testing, is the people who participate. The success of NOCTI is a graphic example of the old saying that "nothing is any better than the people involved in it." If the people, civilian and military, involved in the DANTES-sponsored pilot project this year perform with the same dedication and zeal as have other pioneers in competency testing, we will get verifications and results hoped for and needed to maximize the use of competency testing in civilian education and training. I am optimistic and predict that next year the Military Testing Association will request a report on the success of the project.

Occupational Analysis - A Brief Introduction

B. Michael Berger
U. S. Army Military Personnel Center

At last years MTA, the suggestion was made that the topic of occupational task analysis be considered for the 1975 gathering. This suggestion led to approval of a series of symposia on occupational task analysis, application of data to training and evaluation, and future developments and applications of the processes.

The traditional occupational data collection method used in the armed services as well as in the related civilian communities, was, and in many cases still is, the on-site occupational interview. Some seventeen years ago, the US Air Force began an extensive program to seek new ways of collecting and evaluating such data. The systems which emerged, first the occupational research inventory or questionnaire, and eventually the analytical process known as CODAP (Comprehensive Occupational Data Analysis Programs), have now been adopted by the US Armed Forces and several allied forces such as the Canadians, the Royal Australian Air Force, and the Royal Navy. Similar techniques are being considered or are in use by several civilian governmental and non-governmental agencies.

The three symposia and the scheduled informal discussion period which will be presented here will introduce and discuss in depth the occupational research and analysis processes from several different perspectives. This morning, Dr. Ray Christal will introduce a new series of CODAP programs developed to provide information on that elusive and difficult to define element known as criticality. Ray will be followed by Major Tom O'Connor of the Air Force Occupational Measurement Center who will discuss a model for evaluating training in terms of field use of the material provided in school. Tom will be followed by Commander Colin Nicholl, Royal Navy, our distinguished colleague from across "the pond" as he describes it, who will provide insight as to their use of occupational data to assist in training design.

This afternoon, Mike Strub of the Army Research Institute, will chair the second symposium which will undertake to discuss on-going or theoretical programs for applying collected data to problem solving or decision making processes. During the informal discussion period in the late afternoon, Helmut Hawkins from our office will lead us in a "search for valid measures of criticality" which will set our teeth for Thursday mornings session which he will also chair wherein we will hear of the advances made by the Air Force Human Resources Laboratory toward solving the "problem of criticality" - these tied directly to Dr. Christal's presentation this morning. We will also hear of use of task analysis data in developing criterion reference tests and finally learn of the meaning of that delightful topic, "The Ultimate Catastrophe".

This first MTA effort toward discussion of occupational task analysis and its uses should, we hope, prove interesting and informative. I speak for myself and those who will come before you to encourage your involvement in the program, your questions, criticism, and opinions of the thoughts and ideas presented. We wish an open forum for your ideas. Without further delay, I will now present our first speaker.

The Use of CODAP to Define an Emergency Rating

LT(JG) Allen R. Thuring, USCG

CODAP has been with us for many years now. It has proved to be a useful, many faceted tool in conducting Job Task Analysis studies in the armed forces. In particular, it has been very effective in identifying goals for training and testing programs dealing with often highly technical fields. However, its use has generally been associated with occupational areas which could be considered fairly well structured.

Let me define what I mean by "well structured". These are occupational areas which usually have a clear, precise definition - somebody "knows" what the job entails, even if the definition is faulty. Second, there is an established, trained worker pool. This pool would be an MOS area for the Army, or a Rating in the Coast Guard or Navy. Third, personnel in more senior positions have more experience in the job, and occupy positions of authority over junior men as a result of this experience. Fourth, most if not all aspects of the job are observable by an outsider through examination of routine performance. Fifth, while any job exhibits local peculiarities, the core of the position does not vary significantly from one locale to another. Finally, all trainees in the job subsequently spend a significant amount of time performing skills learned for the job.

This paper deals with the use of CODAP for an emergency rating, Port Securityman in the Coast Guard Reserve. Using the above definition of "well structured", many of the stated characteristics did not apply. First, the official definition of the rating was both outdated and so broad as to be meaningless. Second, most personnel performing tasks associated with this rating have no formal training in this occupational area. Third, positions of authority and responsibility are assigned on the basis of time in service and seniority, but these often are not related to actual experience in the occupational area. Thus the situation occurs where a senior enlisted man with no experience in the occupational field is overseeing the efforts of a very junior man who has more actual field experience and "know how". Finally, most of those men trained in this occupational area do not perform these tasks immediately after training, but must instead retain these skills for possible future use.

To the above conditions, a second major factor in this study was added. The CODAP data was collected from many ratings of active duty personnel who worked in Port Security functions. These people were Boatswains Mates, Gunner's Mates, Yeomen, etc. However, this data was applied to a different group: Reserve personnel on inactive duty in only one rating, Port Securityman, who perform these functions only when activated for national emergencies. For all intents and purposes, the two groups were mutually exclusive. As I'm sure you realize, this application of CODAP is somewhat unusual, and provides new data on the versatility of CODAP in effecting occupational analysis.

The rating in question, Port Securityman, does have a formal definition of sorts. As defined in the Enlisted Qualifications Manual, CG-311, "Port Securitymen are specialists fully qualified in the knowledge, techniques, skills, and equipment of enforcing the regulations and orders relative to the protection and security of vessels, harbors, ports, and waterfront facilities. They enforce the regulations 'Explosives or Other Dangerous Articles on Board Vessels' and 'Rules and Regulations for Military Explosives and Hazardous Munitions'; supervise and control the safe handling, transportation, stowage, and storage of explosives and other dangerous cargoes; in restricted areas, prevent unauthorized persons from entering upon the vessel or waterfront facility, and carefully examine all authorized persons permitted to do so. They are well versed in the field of fire inspection, prevention, detection, control, and extinguishment, and assist in combating fires on board vessels and at waterfront facilities." In addition to this definition, Port Securitymen are also involved with Marine Environmental Protection, especially oil spill prevention and cleanup. However, the scope of this job far exceeds the above definition.

The history of Port Security in the Coast Guard goes back nearly 50 years. However, the current concepts of the job were generally formulated during WW II and the Korean Conflict. At that time, the definition given was quite accurate. The job was performed by reservists called to active duty, while the regular Coast Guardsmen manned ships employed in convoy and invasion duties. With the advent of the Cold War, however, it was determined that these functions should continue. Accordingly, the deactivated Reservists were replaced by active duty Regulars. But there was no Port Securityman rating in the Regular Coast Guard - there still isn't. It is exclusively a Reserve rating. Thus Boatswains Mates, Gunners Mates, Damage Controlmen, and numerous other ratings were used to fill the necessary billets at the peacetime units. Most of these men did so without formal training. What training there was existed on a local level, with OJT by the unit officers the primary source. With a situation such as this, it naturally followed that the most proficient personnel in these duties were those at the unit who had been there the longest, not necessarily those who were most senior. Personnel were not necessarily transferred to the unit on the basis of prior port security experience, and were transferred elsewhere when more pressing needs arose in their own rating. Due to significant differences in the tasks performed, often determined by the size of the port involved, the job when viewed on a national level varied considerably from one locale to another. Finally, the training program was geared to Reservists, who did not perform the tasks on a routine basis, but were expected to retain the knowledge for possible future mobilization. Very few active duty Regulars received this training.

This peacetime Port Security operational area was very similar to that which had been done by Reservists. The Port Securityman Rating Qualifications provided a correct view of the scope of the job. In addition, the training afforded Reservists in the Port Security field adequately reflected the job as performed in the operational units.

This relationship between Reservist training and the actual job performed did not last, however. As time passed, many new jobs and duties were added to the "Port Security" function. These included regulation of dangerous cargo on merchant vessels, monitoring the operation of tank ships and tank barges carrying petroleum and chemicals in bulk, and marine environmental protection, especially in response to oil spills. For all these changes in the operational area, there were no subsequent changes in the written requirements for the Port Securityman rating. Thus the rating requirements, and also the training programs, became increasingly out of date.

A critical juncture came in the early 1970's. With the advent of significantly increased augmentation by reservists of regular units, the Coast Guard Reserve once again found itself doing the job, not just preparing for it. Although the actual performance was only on a part time basis, primarily weekends, it was soon clear that the rating qualifications had to be updated.

The key question was how? The traditional approach was to add new items, thus "fleshing out" the old vehicle. However, there were now operational areas that were not even conceived of when the existing qualifications were written. Rather, the Coast Guard decided to employ CODAP. There would be some changes however. First, the study would be geared around the operational unit, termed the 'Captain of the Port' office. A COTP, as it is called, is centered on a port, includes some surrounding territory on the coastline, and is the prime command in a given area for enforcing federal regulations entrusted to the Coast Guard. Its personnel perform most if not all the 'Port Security' functions within that area. At each Captain of the Port Office, the CODAP questionnaire would be given to all regular Coast Guardsmen performing 'Port Security" functions, irregardless of rating or paygrade. It was up to the individual command to determine who these men were. Thus, each command received an adequate number of questionnaires, decided on a local basis who was to complete them, and then returned the completed booklets to Coast Guard Headquarters in Washington, D.C. With this material, it became possible to define in terms of specific tasks what a COTP office does, and what the job parameters are for all personnel concerned. Differences between locales could be analyzed, as could tasks performed by different paygrades. The data drawn from active duty Regulars could then be used to update the Port Securityman rating qualifications for Reservists.

The development of the new Port Securityman rating qualifications involved several steps. First, the Group Summary Program was used to analyze and compare job data across paygrade. Then the job data was overlapped and grouped in time and task similarity matrices. Analysis of the groups identified in these matrices was then performed. It became clear that pay level was much more important an indicator of tasks performed than rating at an operational Captain of the Port Office. This analysis also made it possible to identify duties limited to one unit or or area, and thus not universal. These duties then could be isolated. The end result of this process was a listing for each paygrade of those tasks performed.

The next step involved taking these lists of tasks and writing qualifications and requirements for advancement that reflected the skills and knowledge needed. This step did not limit itself to CODAP data, however. This was the point where subject matter experts and program managers lent input. Tasks were added which, though performed rarely, had serious consequences if not performed correctly. Some of these were common to other ratings, such as First Aid or firefighting. Others dealt with national security programs which would be assigned the Coast Guard in national emergencies.

These preliminary qualifications were then reviewed by those other headquarters offices which dealt with Captain of the Port Offices. The Office of Marine Environment and Systems had two divisions which were involved. Marine Environmental Protection dealt with Oil Pollution, Pollution Prevention, and Ocean Dumping. Law Enforcement was concerned with dangerous cargo on docks and vessels; fires and explosions in port areas; and vessel surveillance and boarding. The Hazardous Materials and Cargo division of the Office of Merchant Marine Safety was concerned with the handling on both ships and facilities of certain highly toxic/flammable/explosive cargoes. The Military Readiness division checked those areas affected by mobilization requirements and provided input for DOD requirements assigned the Coast Guard. Finally, the various divisions in the Office of Reserve went over the preliminary qualifications. These were our people, and we handle their recruiting, training, and advancement. All these areas had to be evaluated in light of the new qualifications.

Headquarters review completed, the first draft was then sent to the Ports and Waterways Safety School, Yorktown Va. There it was reviewed for completeness and technical accuracy. In addition, there was input by the instructors, all of whom had served in the field performing these duties. They developed an organization along technical lines for the new qualifications, with sections devoted to Break-Bulk (Package) Facilities, Bulk Oil and Chemicals, Tank Ships and Barges, etc.

The draft was returned to headquarters, and reviewed once again by the various program managers. It is interesting to note that in the space of only three months some items had to be changed, due to new Federal Regulations. The final draft was completed in April, 1975, after two years in development. This represented the first complete revision of the Port Securityman Rating qualifications in nearly ten years.

The importance of the new Port Securityman Rating Qualifactions cannot be underestimated. These now establish the standards for advancement of enlisted personnel in the rating. As such they directly affect over 2,200 men. They have input in the correspondence courses, and provide the basis for questions on the Service Wide Examinations. They also become the training goals for the Office of Reserve. Thus they have impact at formal schools, and also on training at the local unit level. These new requirements now allow us to train Reservists to really perform the duties required by the job. This leads to a closer relationship between the Reserve and the Regular Coast Guard in two areas. One is the increased acceptance of individual Reservists by individual Regulars in the field as men who can 'do the job'. It also provides a real benefit to under-manned Regular units where Reservists can come in on weekend augmentation and take over, giving the Regulars a well needed break.

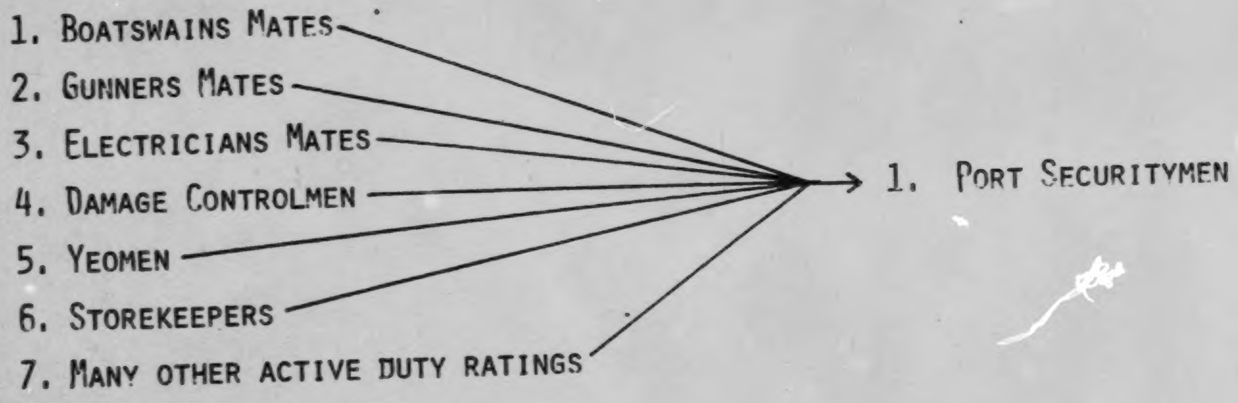
FIGURE 1

A WELL STRUCTURED OCCUPATIONAL AREA

1. CLEAR DEFINITION.
2. TRAINED WORKER POOL.
3. SENIOR PERSONNEL ARE MORE EXPERIENCED IN THE OCCUPATIONAL AREA.
4. JOB TASKS MAY BE INDEPENDENTLY OBSERVED.
5. CRITICAL TASKS DO NOT VARY WITH JOB LOCATION.
6. TRAINEES SUBSEQUENTLY PERFORM TASKS IN WORK SETTING.

FIGURE 2

APPLICATION OF CODAP DATA



GROUP 1:

Active Duty Personnel Performing
Port Security Functions.

1. Boatswains Mates
2. Gunners Mates
3. Electricians Mates
4. Damage Controlmen
5. Yeomen
6. Storekeepers
7. Many other active duty ratings

GROUP 2:

Inactive Reservists Who
Train to Assume These Duties.

1. Port Securitymen

FIGURE 3

THE GROWTH OF THE PORT SECURITY OCCUPATIONAL AREA

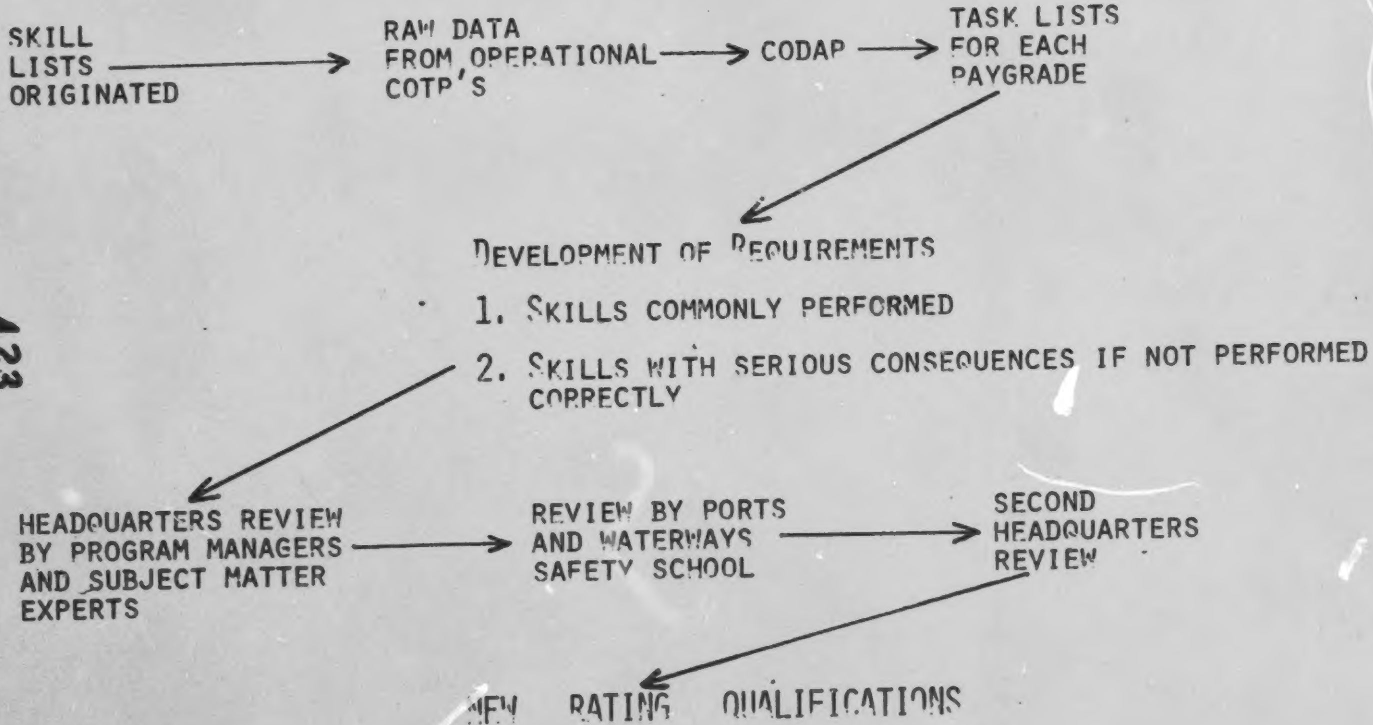
ORIGINAL DUTIES

1. Security of Vessels, Harbors, Ports, and Waterfront Facilities.
2. Explosives Handling.
3. Fire Prevention, Detection, and Fire Fighting in Harbor Areas.

PRESENT DUTIES

1. All Original Duties.
2. Hazardous Cargo Handling on Merchant Freighters.
3. Tank Ship and Tank Barge Regulation.
4. Regulation of Waterfront Facilities Handling Oils and Chemicals in Bulk.
5. Hazardous Chemical Discharge Response.
6. Oil Spill Investigation, Response, and Cleanup.
7. Monitor of Industrial Dumping of Chemical Wastes at Sea.

THE DEVELOPMENT OF THE NEW PORT SECURITYMAN:
RATING QUALIFICATIONS



SKILL LISTS ORIGINATED → RAW DATA FROM OPERATIONAL COTP'S → CODAP → TASK LISTS FOR EACH PAYGRADE

DEVELOPMENT OF REQUIREMENTS

1. SKILLS COMMONLY PERFORMED
2. SKILLS WITH SERIOUS CONSEQUENCES IF NOT PERFORMED CORRECTLY

HEADQUARTERS REVIEW BY PROGRAM MANAGERS AND SUBJECT MATTER EXPERTS → REVIEW BY PORTS AND WATERWAYS SAFETY SCHOOL → SECOND HEADQUARTERS REVIEW

NEW RATING QUALIFICATIONS

Development of Procedures for Evaluating the U S Coast Guard's Post-Graduate and Post-Commission Training Program

Richard S. Lanterman
Psychological Research Branch
Headquarters, U S Coast Guard

J. W. Cunningham
Center for Occupational Education
North Carolina State University at Raleigh

Abstract: The Coast Guard is developing, under contract with the Center for Occupational Education, tools to evaluate their post-graduate and post-commission training programs.

The technique uses a job-task inventory of the entire officer corps. The items in the inventory are educational units instead of the conventional tasks. The educational units represent the entire professional educational requirements of the Coast Guard officer. Each respondent evaluates the inventory items on two scales. On the first scale he rates the amount of the skill or knowledge represented by each educational unit which is required by his present billet (assignment, job). On the second scale he rates the amount of the knowledge or skills which he presently possesses.

Applications of the CODAP analyses of the resultant data to the training management process are discussed.

Background : Each year, a large number of Coast Guard officers and warrant officers go back to school to improve their performance potential or to become qualified for a wide range of billets. Their training can range from a postgraduate degree at the university level to job-specific training from industry. It may be specific to problem areas such as oil pollution or general in scope such as management seminars. The training is intended to enhance performance for specific jobs as well as to develop in an officer the flexibility for assignment to a number of jobs. In short, the Coast Guard has a variety of training programs to satisfy a wide range of training requirements, all with the objective of improving performance or for qualifying and licensing purposes.

For a number of reasons, the Coast Guard has increasingly favored policies which would consider training as a continuous and on-going process that would supplement on-the-job experience.

Recently, the Coast Guard completed a study of the

Academy curriculum. The results indicated that it was neither feasible nor desirable to increase the amount of training at the Academy or the Officer Candidate School precommission levels (though the cost of training, at this point, is less than at a later time). The study also indicated that some precommission training presented at the Academy level should be shifted to post commission training.

Performance considerations, independent of job relevance, which must be considered in evaluating training requirements involve the expectations and motivations of the officer corps or those for whom or with whom they work. Education provides a method of self fulfillment; it enhances personal qualifications, and in general, is considered by the officer corps as one of the benefits of being in the service. A recent officer survey indicates that the majority of officers with a baccalaureate degree desire and expect to obtain a post-graduate degree while in the service. This overwhelming expectation to obtain advanced degrees is probably based on the belief that there is a substantial personal benefit to be achieved from such training. If it is primarily preparation for assignments that is important, then training need not necessarily be as oriented to degree programs as it has been in the past. Also, such training could occur when needed over the period of a career. In any event, the Coast Guard needs a better understanding of officer and warrant officer motivations regarding postcommissioning and post-graduate training.

Assignment procedures try to match officer grade level with billet grade level. However, it is important to note that military grade is vested in the individual and is not determined by the grade of the billet to which assigned. This combined with the service policy of entry only at the lowest grade level, supports the need for degree programs which enhance the individual's status and abilities. Most officers will rotate through a number of billets both in and out of specialty. Rotational patterns vary over time by specialty, by grade, and by individual. Rotational policies, in general, assume a broad performance potential for officers rather than more narrow, specific, and technical performance factors associated with specific billets. However, this rotational policy frequently increases the need for billet-specific training presented over the period of a career. Recently the Coast Guard has civilianized some highly technical officer billets thereby reducing the postcommission training requirements. Continued civilianization, where feasible and desirable, is planned.

Expanding technology, new knowledges, and new roles and missions for the Coast Guard are factors responsible for an accelerating increase in requests for additional training. At the same time that requirements for new training are increasing, costs are also rising. The cost of training will require an increasingly greater share of available resources unless the training can either be provided more effectively or the lowest priority training is dropped each year. Either method of obtaining a solution reouires a thorough study of training requirements.

One method of determining training requirements would be to conduct a job analysis of officer and warrant officer billets and then determine the training requirements needed to support performance in the billet. This method is difficult and expensive. Another method is to rely on the expert judgement of the job supervisor supplemented with a review process. This essentially is the method currently in use in the Coast Guard and it has, in general, been found to work adequately in establishing new training requirements. The current procedure is informal and not systematically implemented by directive. It does not adequately review training requirements previously established. Another potential problem stems from the fact that supervisors tend to express training requirements in terms of conventional educational programs. For example, for a billet which might require knowledge in the statistical and computer sciences, the supervisor would most likely think in terms of a one-year post-graduate degree program in statistics or computer science. However, the billet might only require 15 hours of training in these subjects and some of this training might even be at the undergraduate level. The requirement, therefore, may be overstated with current procedures.

Job-task analysis procedures currently employed throughout the Department of Defense and by the Coast Guard have demonstrated that incumbents accurately describe their own jobs if questionnaires are comprehensive and properly administered. It appears feasible to utilize a questionnaire approach to determine training requirements by replacing job-tasks with education/training 'courses' or 'knowledge unit' type of items. A comprehensive list of such training 'units' or 'courses' could be formatted into a training requirements questionnaire for use by each officer in describing his current billet and also to describe his own personal level of educational achievement. This questionnaire also could include items to determine officers' attitudes and opinions regarding post-graduate training and degree programs.

A major advantage of using a job-task analysis format is the availability of the CODAP package of statistical/computer programs for use in data reduction and evaluation. This package of programs, geared to job-task analysis, far surpasses any readily available programs on the market for similar purposes. The package includes a hierarchical cluster analysis program which might be most useful in clustering training units into courses or programs.

The general study objective then is to develop instruments and procedures which will permit the Coast Guard to evaluate periodically its post-graduate and postcommission program and achieve viable, dynamic solutions to the on-going problems of providing training most relevant to service needs.

Data Collection: One of the most important aspects of this study is the development of a questionnaire which will provide a variety of data outputs of use in **evaluating** contending requirements. In order to achieve this objective, the questionnaire will obtain opinion and attitude data, biographical data,billet training requirements and personal training resources. Each respondent will first evaluate his present billet in terms of courses or training modules required. Educational units will be rated on a seven-point significance (importance to the job) scale. Then he will describe his own educational achievement in terms of the comprehensive training list developed to describe the billet training requirements. Each educational unit selected as describing his educational achievement will be rated on a "level of readiness to use" scale. Thus, each training item in the questionnaire will be considered by the respondent and rated on two scales:

1. If the item is a part of his job requirements he is to rate it on the "significance to his job" scale.

2. An item, whether part of the respondent's job requirements or not, shall be rated by the respondent to indicate his personal educational achievement and also his ability to use the information or knowledge at the present time. Thus, the incumbent will first rate his job, then himself.

This is where we are now. North Carolina State is currently developing the list of educational units which will be the basis for the inventory. This is mainly a matter of collecting expert opinion from program managers in the Coast Guard and consultant educators at universities and training establishments which might be involved in the post-graduate postcommission training process. When the inventory is developed, it will be administered to each officer in the Coast Guard.

Analysis of the results shoud give us, on the one hand, an inventory of the training requirements for every officer billet in the service; and on the other, an inventory of the resources available in our officer corps. Comparison of the two will obviously be an extremely valuable tool for the Training and Education Division who are sponsoring this research.

Criterion-Referenced Personnel Assessment: An Overview

Lawrence S. Buck
U. S. Civil Service Commission

Introduction

The criterion-referenced testing movement is rapidly receiving recognition as an important development in testing. A by-product of this recognition has been ambiguous, nonstandardized concepts and terms seemingly inherent in any newly discovered area of inquiry. A variety of definitions have been offered as well as a variety of procedures for developing criterion-referenced tests (CRTs). To date there is no comprehensive theory of criterion-referenced testing such as exists for norm-referenced testing.

The intent of this paper is to tie some of the loose threads together in a synthesis and summary of the current status of CRTs. This is done in the context of personnel assessment, with particular reference to the military. The **basis** for this report was an extensive literature review. The following topics will be covered: definitions and examples of CRTs; the utility of a criterion-referenced testing approach for the military; types of tests usable as CRTs; guidelines for test and item development; establishing standards; and the application of a criterion-referenced testing approach to the development, implementation, and evaluation of training programs.

Definitions and examples of CRTs

One of the most quoted definitions of criterion-referenced testing is that proposed by Glaser (1971) who defines a CRT as, "one that is deliberately constructed to yield measurements that are directly interpretable in terms of specified performance standards" (p. 41). Popham and Husek (1969) define a CRT as one, "used to identify an individual's status with respect to an established standard of performance" (p. 1). Jackson (1970) suggests that the term "criterion-referenced" should apply only to a test designed and constructed in a manner such that explicit rules linking patterns of test performance to behavioral referents are produced by an objectively-defined process.

A common element running through most definitions is that CRTs reflect an absolute standard of quality and are interpretable in their own right without reference to the performance of other test takers. This contrasts with norm-referenced tests (NRTs) which reflect relative standards and are meaningful only in relation to the performance of others. With a criterion-referenced measure, attention is directed toward each individual examinee and his or her performance relative to a standard. Despite the differences between CRTs and NRTs it is not always possible to tell one from the other with a cursory examination; one must determine how the test is developed, for what purpose, and how the test results are to be interpreted.

The word "criterion" in the term "criterion-referenced tests" is perhaps an unfortunate word as it is easily confused with the use of this word in the term "criterion-related validity." To get around this confusion many authors have suggested other terms for tests which essentially incorporate the basic principles and concepts of CRTs. Some of these are: domain-referenced tests (Millman, 1974); content-standard tests (Ebel, 1971); universe-defined tests (Osburn, 1968); and, objective-based tests (Dahl, 1971). In this paper the terms criterion-referenced test and criterion-referenced measure will be used interchangeably to represent certain test construction rules and methods, and the allocation of competitors to mastery states, i.e., classifying a competitor as a master or non-master depending on his or her performance relative to established performance levels.

It should be emphasized that the word "criterion" in the sense of CRTs does not refer to a criterion in the sense of a normative standard, but to the minimal acceptable level of functioning that an individual must achieve in order to be assigned to a mastery state or level on each domain (or task) included in the test (Millman, 1974, p. 312; Hambleton, Swaminathan, Algina, & Coulson, 1975, p. 6).

Although the popularity of CRTs is relatively recent, you have undoubtedly come in contact with them. Many performance tests, especially in the military, are essentially CRTs. For example, a typing test with an established standard such as 40 words-per-minute represents a CRT if the examinee is assessed solely on a mastery-non-mastery basis where comparisons with others are irrelevant. The Senior Red Cross lifesaving test provides another example of a CRT. Certain performance standards are established--e.g., demonstrate swimming proficiency using a variety of strokes over a specified distance--and the individual must demonstrate the ability to perform at the specified standard. There is no limit as to how many people can possess a skill or skills, i.e., there are no limits on the number of people who can be classified as masters.

While the examples cited are basically performance tests, CRTs can consist of tests with knowledge-type items. CRTs can also be assessed in a variety of ways including paper-and-pencil tests, simulated performance tests, actual performance tests, etc. Again the determination of whether a test is a CRT or a NRT depends on how and for what purpose the test is developed and how the results are interpreted.

Use of CRTs in the military

Criterion-referenced tests appear to be well suited to the assessment of military personnel as they are designed to assess job knowledges and skills in a realistic manner. They can be used as screening devices, diagnostic aids, for evaluating instruction, and quality control of training programs. A criterion-referenced approach would seem to be especially applicable to the development and implementation of training programs. This use of CRTs will be discussed more thoroughly later in the paper. Swezey, Pearlstein, and Ton (1974a) report on an effort by the Army Wide Training Support Group of the Air Defense school at Fort Bliss, Texas to include performance tests in correspondence course evaluation. Some of the correspondence courses "contain an OJT [on-the-job training] supplement which is evaluated via a [criterion-referenced] performance test administered by a competent monitor in the field where the correspondent is working" (p. 1-44).

Swezey et al. (1974a) also surveyed a number of army bases to determine the prevalence of the use of CRTs. Some of their general conclusions are: the use of CRTs in the army is definitely increasing; 88% of the interviewees felt that CRTs should receive high or top priority in terms of army assessment programs; 60% felt that CRTs should replace most or all NRTs; many felt that criterion-referenced testing would be useful and practical for measuring areas other than job performance skills; over two-thirds of the interviewees indicated that increased expense may be a problem in the development and use of CRTs. In general it was held that more soldiers are being evaluated more effectively through the application of criterion-referenced testing in a diversity of areas including field wiring, food services, basic combat training, and advanced individual training. There was a great deal of respect among the interviewees for the utility and practicality of criterion-referenced testing.

MOS (military occupational specialty) testing in the army is primarily norm-referenced testing. The present move to performance-based testing in the army has led to a consideration of the use of CRTs for MOS testing, although to date there has not been a great deal of research in this area. CRTs, either in conjunction with NRTs (perhaps as screening devices) or alone may be adaptable to MOS testing. Where quotas for advancement are established, cut-off scores may be set to limit the number of eligibles to coincide with the quotas. If ranking of individuals is considered a necessity for MOS testing then CRTs alone would not be applicable. One solution may be to rely on other factors such as supervisors ratings, credits for awards, time in service, etc. to rank those who have reached the mastery level on a CRT.*

---

*This brief discussion of MOS testing is based on a personal communication with Dr. Raymond O. Waldkoetter, Jr., former Technical Advisor at USAEEC.

Types of tests usable as CRTs

There are a variety of types of tests that can be developed for use as CRTs. It is very important to keep in mind that the type of test selected is situationally specific. Job knowledge tests might be appropriate in one situation, performance tests in another, and combinations of the two in other situations. The test must be relevant to the job situation and to job performance. Test development costs and other considerations such as ease of standardization, objectivity, practicality, and reliability will help to determine the type of test for a particular situation.

If job knowledge tests are used as CRTs one must ensure that they measure only necessary and relevant aspects of job performance. Vineburg and Taylor (cited in Swezey et al., 1974a) suggest that, "job-knowledge tests are valid for measuring proficiency in jobs where: (1) skill components are minimal, and, (2) job knowledge tests are carefully constructed to ensure only that information that is directly relevant to performing the job at hand" (p. 1-20).

Job performance tests such as actual work-sample tests are especially applicable for use as criterion-referenced measures although they are usually very costly to develop and administer. When used as a CRT, a performance test must mirror realistic job requirements and must measure the ability to perform critical tasks. The job situation must be simulated in a controlled manner so as to effect a standardized environment in which performance can be readily and reliability assessed.

Job-component techniques and subjectively-scored techniques are other examples of performance tests. Job-component techniques simulate the requirements of a specific subtask of the job, testing limited, well-defined portions of overall job performance. A job would be analyzed into components and the performance on each of the components evaluated.

Although a carefully constructed performance test would have content face validity, there are a number of problems associated with performance tests. If a test utilizes complex equipment it may be difficult to maintain the equipment in a constant condition for all candidates. The environmental conditions may vary. Standardization becomes difficult.

With some objectively scored work-sample techniques using actual or simulated on-the-job environments, the applicant performs the same work--at entry level standards--as that involved on the job. The usefulness of this type of CRT will depend on how well the important aspects of job performance are represented in the test situation. If most applicants would be able to easily learn these skills on the job in a reasonable amount of time, these skills should not be tested for; selection should be made on some other basis.

Guidelines for Test and Item Development

An essential step in the construction of CRTs is the development of a comprehensive test plan. Swezey et al. (1974b) discuss a number of issues that must be taken into account in the process of developing the test plan; e.g., costs, manpower availability, time availability, and equipment and facility availability. These issues affect any determination of the methods used to test objectives. The next step is to plan the type of item (item format) and the level of fidelity (realism) (p. 3-1). Finally, a decision must be made on how many items should be included in the test. The test plan should be documented.

It is essential that the performance domain measured by CRTs be defined in terms of observable behaviors and that the test be a representative sample from which competence can be inferred. This can best be accomplished by the development of test items and test parts based directly on a useful, well-defined job analysis; i.e., one which insures that the tasks and/or knowledges, skills, abilities, or other worker characteristics (KSAOs) compiled are representative of the behavioral domain of interest and thus insures content validity.

Job analysis

The purpose of a job analysis is to provide descriptions of the actual behaviors and/or KSAOs necessary for successful job performance and their relative frequencies and importance. Rundquist (1967) defines the objectives of a job analysis in terms of three elements: "The bebehavior the person is expected to demonstrate; the conditions under which he or she is expected to demonstrate the behavior; and, the standards at which the behavior is to be demonstrated" (p. 2). Failure to conduct an adequate job analysis can result in false inferences about the KSAOs necessary for successful job performance. In planning a CRT one should: identify the situations and conditions under which relevant tasks or KSAOs will be used; identify the tasks or KSAOs necessary for successful tasks or KSAOs. A distinction must be made between those skills which are normally developed and/or improved on the job, and those which are basic to learning the job. In determining whether a skill is necessary for job entry one should determine, as far as possible, ff learning the skill(s) would take considerable time, and the degree of difficulty most people would have learning the specified skill(s).

A major problem in a job analysis is designing task statements that clearly and unambiguously represent the relevant domain for test development purposes. For each task it is necessary to decide what level of proficiency each competitor should demonstrate. It is essential that each task description include criteria or performance which competitors must meet.

It is crucial that the test constructor be alert to some common biases when conducting a job analysis. Examples of such biases are: ommission of pertinent requirements from the actual criterion; unequal scale units; inclusion of extraneous job requirements; and, improper weighting in combining aspects of the actual criteria. A careful, thorough, and systematic job analysis can for the most part dispense with these biases.

Item development, selection, and analysis

Once the purposes and objectives of the CRT have been defined, the next step is to construct and select test items or test tasks to measure the chosen objectives. It is probable that a great many items could be constructed for each task. Testing constraints however, would likely limit the number of items usable. It will be necessary to draw a sample of items from a population of items. Simple random or stratified random sampling procedures would seem applicable. Millman (1974) states that, "unless items are selected randomly the estimate of a person's level of functioning loses meaning and the interpretability of the test score is reduced" (p. 339). Because of the unique demands of CRTs, it appears necessary to develop new methods for determining how many items are needed to sufficiently sample a given objective.

One must ensure that items selected for use in CRTs are accurate measures of the criterion behavior. The items or test tasks must clearly be members of the relevant domain. The number of items to be used is influenced by such factors as the amount of testing time available, the level of performance required, the type of position being tested for, and the cost of making interpretation errors. It is essential that the test be of a reasonable enough length to allow generalizations about which specific tasks an individual can or cannot do. Minimum levels of performance must be specified describing the least amount of competence the eligible competitor is expected to have. The items should be created and constructed according to specifications established in the test plan.

One method that seems appropriate for constructing and analyzing items for CRTs involves the use of a panel of subject matter experts to decide how to test for different tasks. The panel could decide which items would best assess performance for each task.

Osburn (1968) suggests the use of the "item form" which is a complete set of rules for generating a domain of test items which can be used to measure a particular objective accurately. Osburn considers item forms important in that they provide a "basis for generation...to a universe of content...contained in the operational definition of the procedures used in generating and sampling items that go to make up the test" (p. 96). Item forms can be a valuable aid to validity determination in that, "what the test is measuring is operationally defined by the universe of content as embodied in the item generating rules" (p. 101).

Osburn lists the following characteristics of an item form:

1. It generates items with a fixed syntactical structure.
2. It contains one or more variable elements.
3. It defines a class of item sentences by specifying the replacement sets for the variable elements (p. 97).

Although the item form approach has some possibilities the use of this approach may be quite limited. As Hambleton et al. (1975) point out, "it does not appear likely that the notion of specifying content via the use of item generation rules will be applicable to many subject areas" (p. 15). Millman (1974) feels that "item forms may not adequately capture the variety of test tasks an item writer may wish to employ in measuring a knowledge, skill, or attitude" (pp. 331-332).

The traditional or classical methods used to select and analyze items for NRTs are generally inappropriate for selecting and analyzing items for CRTs. Coulson and Hambleton (1974) contend that "selecting items for domain-referenced tests on the basis of their statistical properties would destroy the requirement that the items are representative of the domain of items" (p. 14). Many NRTs are intended to provide valid discriminations among people at all levels of achievement and a spread in item difficulty as desirable. This is not the case for CRTs which are not intentionally constructed to discriminate among people at all levels, i.e., variability is not maximized for CRTs. Item discrimination indexes based on the biserial, pointbiserial, and tetrachoric correlation coefficients normally used with NRTs are inadequate for CRTs because of their dependence on variability and a normal distribution.

Technically speaking, item difficulty indices such as proportion correct indices (p values) used in the construction and selection of items for NRTs are an irrelevant concept with regard to CRTs. Either a CRT item measures the criterion or it does not and interest is not primarily centered around the proportion of people who do not get an item correct. Pragmatically speaking, however, classical indices may be of some assistance in detecting flawed items.

CRT items do not necessarily have to discriminate to be included in a CRT. Negatively differentiating items--those answered more often by the less knowledgeable than by the more knowledgeable--may serve as indications of bad items and such items should be carefully scrutinized to determine their worth. However, items should be eliminated only if they do not conform to the domain definition. Items are not automatically discarded because all competitors answer them correctly.

The methods of item analysis suggested for CRT development are varied and still largely theoretical. As previously indicated, classical item analysis instruments or indices (those commonly used with

NRTs) are either generally inappropriate or they may give spurious estimates of the worth of CRT items. The most pragmatic solution appears to be the use of a panel of experts or independent reviewers to evaluate an item's relevance to a task or criterion, or the success the items have in representing the performance domain from which competence is to be inferred. A panel of experts could also be presented lists of items and list of objectives with instructions to match the items and objectives. A high percentage of correct match ups of items and objectives would be desirable to indicate good items.

Items could also be selected and analyzed by administering the items to two groups of people known to possess different degrees of skill with respect to the objective measured. A problem with this method is the difficulty in determining that the different groups have the same general intellectual abilities or other characteristics that could affect test results.

Item reliability could be assessed by calculating the proportion of people whose item scores are the same on a test-retest basis. Under this procedure examinees should be administered two sets of items drawn from the same domain or the two sets could be intermingled in a single administration.

Establishing standards

Standards refer to the minimally acceptable score level above which mastery of an objective is inferred. Standards might also be defined, "from the 'top down' as the maximum deviation from perfect performance that is still considered demonstration of mastery" (Zieky, 1973, p. 1). If standards are set too low, there is an increase in the probability that non-masters will be classified as masters. Many factors must be taken into consideration when establishing standards, the most crucial of which are the costs of making interpretation errors, i.e., of misclassifying people as masters or non-masters. Swezey et al. (1974b) list the following factors that should be considered when establishing cut-off scores:

1. Immediate manpower needs - if manpower needs are very high it may be justified to lower cut-off levels especially if errors are less critical than no performance at all.
2. Upper feasible scores for an established "master"- a target may be placed so that even the best marksmen may score only 50 percent hits. If we set a cut-off score at 70 percent, we will pass no one at all.
3. Criticality of the objective - the greater the risk of substantial damage to persons or to property, the higher the cut-off score should be (p. 6-12).

The procedures discussed below used to establish cutting scores were extracted from a paper by Zieky (1973). He stresses, above all, that under no circumstances should an arbitrary percentage be picked as a cutting score because that number has been traditionally associated with a passing score. Nor does he advocate the use of blanket scores, i.e., deciding that a cutting score set for certain item sets is appropriate for other item sets with other subjects. Any decisions made concerning the setting of a standard must incorporate some analysis of item and test data.

One approach entails the use of subject-matter experts (SMEs) to examine each item and decide which items minimally qualified competitors should answer correctly. The cutting score then becomes the sum of the items determined to be necessary for a specified level of competence. The SMEs must be aware of the exact purpose for which mastery is to be judged. Angoff (1971) has suggested that this method can be improved by having each judge state the probability that an acceptable competitor would answer each item correctly. The probabilities would then be summed to determine a cutting score.

Zieky (1973) cites another method that could be used, "where at least a partial hierarchy of skills exists such that attainment of terminal objectives is of extremely low probability unless certain enabling objectives have been met. Success at long division, for example is not likely unless the student can subtract and multiply" (p. 4). Pooled judgments from SMEs are compiled based on decisions about whether an item represents behaviors that are required in future work, and a cutting score is set to reflect the necessary behaviors.

Another potential approach involves the use of an expectancy table by which scores on a CRT are related to scores on a predicted variable. Zieky (1973) defines an expectancy table as, "essentially a way of summarizing a bivariate frequency distribution to allow the prediction of one variable given knowledge of another" (p. 7). The cutting score is based on, "the probability of reaching various levels on the predicted variable" (p. 7).

Another possible method would involve the use of a group of subjects considered to be minimally competent. This group would be given the CRT and the cutting score for future administrations established as the mean score of the group. This method tends to insure a realistic cutting score in terms of the percentage of the population that is judged to have reached mastery on the basis of the test. Or, preselected groups could be used, one chosen so that it is likely to possess the required skills and the other group so that it is unlikely to possess the requisite abilities. The CRT would be administered to the two groups and a cutting score selected that best discriminates between the two groups.

A criterion-referenced approach to training programs

The most potential for a criterion-referenced approach in the military would seem to be in the area of training. A criterion-referenced testing approach lends itself well to the specification of the training objectives as well as the assessment of the objectives and the training program. "Criterion tests are designed to mirror each set of enabling objectives and their behavioral components. Since criterion tests derive their instructional worth from their one-to-one relationship with a performance standard, an (individual's) score on a criterion measure indicates exactly his learning proficiency or deficits" (Basic Officer Training Systems Design, 1972, p. 12). Other areas in which CRTs can be of value in training are the evaluation of instruction and individualized instruction.

The overall goal of a training program should not be to rank participants but rather to ensure that participants learn the required skills; that they can perform on the job at acceptable proficiency levels. How much more or how much less competent the person is, is of no interest from a strictly training point of view. A thorough job analysis will determine the requisite skills and knowledges as well as providing a means of assessing them. With a criterion-referenced approach, the training curriculum can be specified in terms of behavioral objectives derived from the job analysis. As Swezey et al. (1974b) indicate, "CRTs are developed from objectives that tell what people must do to successfully complete or perform certain tasks" (p. 2-1).

The task inventory identifies job actions which will become the training actions in learning objectives. The ideal situation is to evolve training tasks that are as close as training conditions permit to the job tasks. If identical job actions cannot be evolved in the training program, it becomes necessary to determine whether transfer takes place. The major reason for a training course should be training in the desired performance standard task, which is best accomplished with a criterion-referenced testing approach. Of primary concern to the training program is the cost of turning out students who are not proficient to the desired degree. Swezey et al. (1974a) point out that, "if the analysis of the job reveals a large number of critical tasks or individual tasks whose criticality is great, then the cost of supplying a non-master can be assessed as high, and great effort is justified in developing a training program featuring high fidelity, costly CRT (criterion-referenced testing") (p. 1-52).

Training program evaluation

A criterion-referenced testing approach is well suited to evaluate a training program. Millman (1974) feels that, "tests referencing the specific domains of learner behavior to which the instruction effort is directed have a much better chance of detecting areas in which the program has been successful or is in need of modification. It is not necessary to throw out an entire program. Improvements can be targeted to the program goals which were not met" (pp. 386-387). Hanson and Berger (cited in Swezey et al., 1974a) in a discussion of quality assurance programs, "conclude by noting that quality assurance, as applied to criterion-referenced programs, would act to ensure that the specified performance levels will be maintained through the life of a program" (p. 1-34).

Absolute rather than relative criteria are essential for quality control. We are more interested in whether the student can demonstrate mastery of course objectives than relative comparisons of a student to other students. CRTs provide absolute criteria that are specific to the task objectives. Properly designed criterion-referenced training tests should indicate where problems exist in the training program and are more likely to reflect instructional problems than testing problems when failure occurs.

Individualized instruction

It is a well known principle that people learn at different rates. Traditional systems with lock-step learning rates do not allow for individual learning abilities. Individualized pacing and instruction allow the persons to learn at their own rate and permit adaptations to individual differences in the experiences students bring to the course. CRTs can be used as pretests to determine where each student should begin his learning on entrance to the training course. A student is tested for mastery of content of each training unit. The results of these tests serve as the criterion for determining whether or not students should proceed. CRTs are the tests best suited to determine whether the student has mastered a training unit or if remedial work is required. It must be reiterated that the primary goal of a training program is to supply individuals who can competently perform on the job, and a criterion-referenced based individualized training program is the best means of achieving this goal.

Conclusion

The criterion-referenced testing movement appears to be gaining great impetus in the military. The adaptability of CRTs to a wide variety of situations and uses enhances their value to the military. The groundwork for a comprehensive theory of criterion-referenced testing is being formulated but much remains to be done. More research is needed in the application of statistical methods to the development of tests and in assessing the final product. More research is needed in the use of CRTs with "soft-skills" where the emphasis is on knowledges and understanding as opposed to specific skills. We need to be able to better analyze the costs associated with misclassifications. Above all we need ways of reducing the extensive costs often associated with CRT development, although in the long run criterion-referenced testing may prove less expensive than norm-referenced testing in terms of insuring the efficient output of well trained personnel.

Among the findings in the survey of army bases by Swezey et al. (1974a) was the need for more rigorous development or application of CRTs. They found that "a major omission in the development of CRTs...is the lack of test evaluation. There was virtually no consideration of test reliability and/or validity" (p. 2-27). Although content validity is generally established for CRTs, procedures for assessing other types of validity and reliability are needed. The favorable and optimistic attitudes of those presently working with CRTs in the army indicates that the future of criterion-referenced testing in the military is bright.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd. ed.). Washington, D. C.: American Council on Education, 1971.

Basic officer training system design. Training Division, Metropolitan Police Department, Washington, D. C.: Author, 1972.

Coulson, D. B., & Hambleton, R. K. On the validation of criterion-referenced tests designed to measure individual mastery. In R. K. Hambleton (Chair), *Recent Developments in Criterion-Referenced Assessment*. Symposium presented at the annual meeting of the American Psychological Association, New Orleans, Louisiana, 1974.

Dahl, T. *Toward an evaluative methodology for criterion-referenced measures: Objective-item congruence*. Paper presented at the annual meeting of the California Educational Research Association, 1971, San Diego, California. (ERIC Document Reproduction Service No. ED 052 253).

Ebel, R. L. Criterion-referenced measurements: Limitations. *School Review*, Feb. 1971, 79(2), 282-288.

Glaser, R. A criterion-referenced test. In W. J. Popham (Ed.), *Criterion-referenced measurement (An introduction)*. New Jersey: Educational Technology Publication, Inc., 1971.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. *Criterion-Referenced Testing and Measurement: A Review of Technical Issues and Developments*. A symposium presented at the annual meeting of the American Educational Research Association, 1975, Washington, D. C.

Jackson, R. Developing criterion-referenced tests. ERIC Clearinghouse on Tests, Measurement, and Evaluation, Princeton, New Jersey: 1970. (ERIC Document Reproduction Service No. ED 041 052).

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education*. Berkeley, California: McCutchan Pub. Corp., 1975.

Osburn, H. G. Item sampling for achievement testing. *Educational and Psychological Measurement*, 1968, 28, 95-104.

Popham, W. J., & Husek, T. A. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, Spring 1969, 6(1), 1-9.

Rundquist, E. A. Job training course design and improvement (2nd. ed.), 1967. (Research Report SRR 71-4), Naval Personnel and Training Research Laboratory, San Diego, California.

Swezey, R. W., Pearlstein, R. B., & Ton, W. H. Criterion-referenced testing: A discussion of theory and of practice in the army. Prepared for Army Research Institute, Unit Training and Educational Technology Systems Area Performance Measurements and Standards Work Unit. Reston, Virginia: Applied Science Associates, Inc., 1974. (a)

Swezey, R. W., Pearlstein, R. B., & Ton, W. H. Developing criterion-referenced tests. Prepared for Army Research Institute for the Behavioral and Social Sciences, Unit Training and Educational Technology Systems Technical Area Unit Training Standards and Evaluation Work Unit. Reston, Virginia: Applied Science Associates, Inc., 1974. (b)

Waldkoetter, R. O. Jr. Personal communication, September 5, 1975.

Zieky, M. J. Methods of setting standards for criterion-referenced item sets; and, Application and adaptations of classical test theory for application to criterion-referenced measures. Princeton, New Jersey: Educational Testing Service, 1973.

EVALUATION OF MANPOWER POLICIES USING COLLECTED OCCUPATIONAL-JOB ANALYSIS INFORMATION

Raymond O. Waldkoetter
U.S. Army Research Institute for the Behavioral
And Social Sciences, Arlington, Va.
and
Peter F. Newton
National Security Agency/Central
Security Agency, Fort Meade, Md.

The advanced methodologies generally used by the Armed Services now to analyze occupational task inventory data have been skillfully developed and apply to well designated areas of occupational research. These methodologies explore everything from the work performed by job incumbents, to what pay should be given for a job, how the job should be designed, through the effects of job satisfaction on individual worker decisions in a job context. Only a few applications of these methodologies are mentioned to give some idea of their utility and the detailed analytical directions which result from basic task inventory data.

In applying one or all of the advanced methodologies, it is interesting to observe that the techniques used are clearly designed to objectively relate the occupational data to an informational value for areas where the research can be profitably interpreted. If questions occur in personnel allocation, training, utilization or retention and are validly conceived, then some methodology or combination of such can be selected to reply to a particular question or questions as presented. When questions about manpower policies are expressed as to what should be done to improve or correct a situation, appropriate answers can be sought by structuring the desired techniques to carry through the most pertinent job analysis and reporting sequence.

Examples of manpower policy questions which are raised are quite often complex and need a little review before settling on what approach and techniques can be pointedly productive. A catalogue of policy questions might be feasibly constructed in the future to relate the question content to the most probable solution using the Comprehensive Occupational Data Analysis Programs (CODAP). It is not intended to go over the CODAP Analysis System, which the Air Force (Christal, 1973) has so creatively developed to answer task-based management questions regarding job incumbents in what may be termed well-defined policy areas. We want to single

---

The views expressed in this paper are those of the authors and do not necessarily reflect the views of the Army Research Institute, Department of the Army or National Security Agency, Central Security Agency.

out certain less-precise management oriented questions to which similar occupational research information may be applied with some adaptation of techniques for processing occupational data. A few questions can serve to describe some of the less-precise categories of manpower policy demands which may befall a researcher and yet offer an abundant exercise of skill and innovation.

Questions to be evaluated, according to policy solutions, based on collected job inventory type data rather than an interview technique may be presented typically in one or more of the following inquires:

1. Do certain types of workers perform at a level of mission efficiency higher or lower than others?

2. Are some job tasks in a given MOS job and skill level beyond the capabilities of the personnel assigned?

3. Are duties/tasks assigned so that equal opportunities for job success can be expected?

4. Will certain duties/tasks increase job satisfaction and can these be reinforced during training or OJT?

A policy solution will be defined as analyzing job data, and interpreting these data in some informational output, to allow a value to be placed on the information and provide a basis for contructing a course of action.

Consensus is urged and received on the above inquiries using command direction and experienced opinion quite frequently. The possibility these inquires are turned away from job content and structure, to managerial viewpoints perceived as exclusive policy questions, follows perhaps a less critical, judgmental solution style. This happens when individuals, interested groups and separate commands are tasked with surveying a manpower policy requirement without enough technical structure. Lacking problem solving structure, an effort is made to find a solution that adheres to somewhat open administrative interpretation. In this way there is ample room to adapt to a general administrative policy guideline, while not being tied down to any detailed policy specification. This is fine so long as precise judgment is not wanted and policy seekers are willing to accept some latitude in individual or command responses.

An old example such as the "zero defects" program might describe a kind of policy requirement where there is supervisory emphasis, but seldom detailed procedures for identifying what should be supervised in the duty/task area to lessen errors, shortfall or whatever would be a measure of higher manpower efficiency. "Zero defects" could, oddly, also be achieved by a drop in production, assuming there was a qualitative improvement. To deal with evaluating a manpower policy so that there are sufficient quantifiable data imposes a need for specifying those duties and tasks which can show how managerial and supervisory policy perceptions and values may be demonstrated and then validated.

WORKER-ORIENTED AND MANAGER/MANPOWER POLICY CONTRASTS

Job incumbents will reflect naturally via the task inventory what they are doing according to the survey instructions. Management and supervisory questions will be asked about what is probably being done and to what extent and by whom in particular situations. This approach may be distinguished as worker-oriented in that task data are generated by job incumbents. Worker-oriented in this sense is different from Gordon and McCormick's (1963) definition where they related worker-oriented to "actions the worker is performing without reference to the job or product involved."

Conversely, when the data collected are reliably useful as occupational information and collected from task estimates of commanders, supervisors and other policy management authorities, the evaluation of manpower policies may be attempted to retain a management based value system. This might be accomplished when management-level personnel evaluate task scales, not based on effects as experienced by workers, but as managers experience judgments in giving task estimates. Where incumbents would yield "time spent" on tasks, one would ask supervisors how much time should be spent to satisfy one or more mission objectives.

Rigorous safeguards in preparing materials for occupational analysis objectives must be designed if a policy, management level perspective is to be explored. It is self defeating to say the least if manpower policy judgments are not controlled to account for any biases that could appear in a policy test regarding some measure of task performance in manpower or personnel programs.

From the earlier sample inquires enumerated above to develop policy-answer interpretations, conditions will be described whereby I think applicable rating data could be generated to supply sound findings. Policy task ratings should be instrumental in providing definition of variables in the solution process.

Inquiry 1

In planning for manpower utilization at the several levels of command, there are situations where levels of enlisted effectiveness and producttivity become an issue of concern. Questions are heard about mission capability and whether performance of given groups of workers produce any differences in how effective types of units might be. Depending on what a definition of mission capability, productivity, or unit effectiveness might be, it should be possible to collect supervisory ratings about whether some incumbents perform tasks at a level of mission efficiency higher or lower than others.

Since the supervisors are charged with maintaining standards of job performance and evaluating the job incumbents, they can likely judge how productive selected workers are on given tasks and the extent to which the workers are utilized. For example, should a question arise so that it is necessary to differentiate workers on the basis of a career-feeder or job-assignment pattern, a task inventory of those tasks most relevant to mission effectiveness could be evolved and supervisory ratings obtained to record to what degree of effectiveness the categorized career feeder groups performed. A scale would have discrete points to parallel the familiar "time spent" scale, adapted to measure "time spent" in doing particular tasks under specified duties. A complex job situation may have noticeable differences which are based on team efficiency or unit behavior in contrast with individual job competence.

Inquiry 2

Following from the question about "level of mission effectiveness" another problem can surface. Commanders/supervisors could well ask if a less-effective group was encountering MOS job tasks at a given skill level going beyond their capabilities. Here the situation would depend on more than factors of aptitude, time in grade, and last school completed. Location of last assignment and the job structure might be handicapping incumbents because a degradation of skill and the span of duties would operate to place some tasks outside the immediate capabilities required.

Worker-oriented task data could actually allow one to conclude in honest error, that more time was "spent" on some "very difficult" tasks by a less-prepared or unqualified group, regardless of time spent by personnel with better aptitudes, more experience and more technical training. The qualitative supervisory judgments under the best statistical rating controls could give a dimension whereby the scaled output of workers becomes a source of comparative distinction. A supervisory trend or a scale about how well some tasks were done, and when they were done in terms of duty sequence, could strongly suggest whether a less effective group could fulfill all essential duties and tasks.

Again preparation of a task-rating scenario, instructions with suitable rating formats, and clear sequential steps would govern what manpower policy guidance would result with adequate reliability and practical validity. Results for this inquiry could become evident in asking also whether selected tasks needed more remedial OJT from supervisors, than the same tasks for other workers.

Inquiry 3

A third question which could demand attention as a manpower policy responsibility is one which causes occasional disruption in scrambling for quantifiable data from subjective policy judgments. This question however is fraught with irritants that make a researcher hesitate before plunging into the problem. When the inquiry is received as an interrogatory for a legal opinion, it can arrive as: Are duties and tasks assigned so that equal opportunities for job success can be expected?

Obviously, the topic "job success" and its meaning depends on numerous variables in the job context. An objective might only relate to delineating that "equal work is expected for equal pay". Supervisory task ratings might then be collected on scales designed to indicate whether tasks are encountered both quantitatively and qualitatively in duty position assignments. Whenever a duty position in an MOS, as indicated by supervisory ratings demonstrated the policy was one of having incumbents exposed to a representative sample of duties and tasks, and a balance "load" of difficult tasks, an argument for a fair administrative allocation of work could proceed.

A cross-check with worker inventory data would be almost mandatory. Supervisors can surely record, however, which things have to be accomplished in a job and those tasks more associated with degrees of job competence. So if one may accept that each incumbent would experience similar amounts of exposure to tasks and job standards, there is a reasonable opportunity for success where workers are competent. This discussion need not refer to racial or sexual bias in jobs, but can come about should an inequity in grade for comparable jobs bear on a timely discussion of compensation practices.

Inquiry 4

A last major question: Will certain duties/tasks increase job satisfaction and can these be reinforced during training or OJT? refers to a managerial manpower policy to identify the tasks making for increased satisfaction in the job situation. Recruitment, retention or efforts to improve productivity are frequently the source of this sort of complex inquiry. Deriving the policy, i.e., knowing what commanders and supervisors could reflect as their principal judgments in defining satisfaction, would produce the data from which satisfaction information is organized to direct the preferred course of action.

The same task inventory also administered to job incumbents could have scales designed along the several dimensions of job satisfaction. Supervisors could rate, for example, which tasks held degrees of interest for workers, those tasks generating favorable work relationships, tasks in terms of difficulty to please supervisors, tasks having value outside the job situation for incumbents, and those tasks which could accelerate promotion potential for subject workers. (Smith, Kendall & Hulin, 1969).

Although there is no substitute for acquiring the worker-oriented responses to tasks about job satisfaction, occasions will prevail whereby policy level collection of supervisor ratings can yield measure s of job satisfaction that may escape worker-level perceptions. Consequently, if there are policy measures of job satisfaction their related tasks are traceable to training and OJT contexts to enhance resultant job satisfaction. Commanders and supervisors could well affect job satisfaction through policy that relies minimally on the immediate worker-task interface. Scales for job satisfaction, clarifying management policy and interaction with jobs, have a contribution that would not usually have a parallel in worker data because of time and experience variables constraining worker observations.

DISCUSSION

Technology in occupational research has been sufficiently developed that its pervasive effects are being seen in nearly all activities having to do with the personnel needs and manpower systems of the Armed Services. Through the initial and current development of occupational analysis methodologies an orientation toward the worker inventory data has largely prevailed it seems, and naturally so. But as questions related to occupational data analysis have occurred in the area of manpower policy, there is a need for moving the emphasis of analysis to a managerial or supervisory evaluation of recognized incumbent performance of job tasks.

Major types of inquires concerning manpower policies may have an identifiable structure, though less precise than the questions formulated in creating answers for occupational information based on a direct collection of incumbent job inventories. To define and evaluate a manpower policy, an essential effort must emerge, I believe, to obtain data which are related to supervisory views of worker performance. Gaining consensus by survey or argued opinion in manpower policy resolution certainly has a place after data are acquired and digested. However, it seems to me that in the areas of manpower and personnel functions, there exist particular issues in the treatment of worker-mission efficiency, the interface of job-structure and worker capabilities, the judicious handling of career opportunities, and the requirement for worker and supervisory job satisfaction, which need a thorough contrast of worker-supervisory data with a resolution through a managerial policy evaluation process (Uhlaner, 1970).

With manpower policy inquiries there are anticipated problems in being certain that rating scale or judgmental procedures are modeled to gather data at a policy level and are based on specific job tasks as verified for selected incumbents. Situations exist in which manpower policy decisions must derive from expert managerial sources to project a given directional force. Strategies for courses-of-action must come from managerial evaluation of previously digested occupational information derived from analysis of worker inventory data.

A trend in occupational research is gradually aiming the methodology and techniques of analysis at greater participation by command and supervisory personnel in my opinion. And, as the techniques for supervisory interpretation and reporting of work data are fully systematized, we shall surely increase the possibility of predictable manpower policy planning and decision making.

REFERENCES

Christal, R.E., The United States Air Force occupational research project. Paper presented at the State-of-the Art Symposium for Occupational Research and Development, Navy Personnel Research and Development Center, San Diego, California, July 1973.

Gordon, G.G., and McCormick, E.J., *The identification, measurement and factor analysis of "Worker-oriented" job variables*. Lafayette, Ind. Purdue University, Occupational Research Center, July 1963. (Prepared for Office of Naval Research under Contract Nonr-1100 (19), Report No. 3).

Smith, P.C., Kendall, L.M., and Hulin, C.L., The *measurement of satisfaction in work and retirement*: *A strategy for the study of attitudes*. Chicago: Rand McNally, 1969.

Uhlaner, J.E., *Human performance, jobs, and systems psychology* -- *The systems measurement bed*. U.S. Army, Washington, D.C.: BESRL Technical Report S-2. October 1970. (AD 716346).

FROM JOB TASK ANALYSIS
TO
BEHAVIORAL OBJECTIVES "BY THE NUMBERS"

By

Raymond D. Bland
Lieutenant Commander, USCG
Chief, Development Division
U. S. Coast Guard Institute
Oklahoma City, Oklahoma

ABSTRACT

The job task analysis provides sound data for training course design. Literature in the field of educational technology provides little if any help to novice instructors in their task of converting that data into training materials. New instructors at the Coast Guard Institute, a correspondence course center, are directed through a methodical, closely supervised process which allows them to use the data without confronting the jargon and abstractions of the educational technologist. At the end of the process, the instructor will have an outline for his text and an outline for the criterion test for the course. An unexpected spin-off of the process is the instructor's understanding of systems development and behavioral objectives, achieved easily by a retrospective view of his completed work.

The job task analysis provides an outstanding source of data for determining training course content. Unfortunately, though, there is an immense chasm between the job task analysis and the final text for the training course. Typically this chasm is glossed over in systems development charts by a single arrow between the Job Task Analysis block and the Behavioral Objectives block. The literature abounds with theories on the nature of those blocks without regard for the existence of the chasm. In fact the chasm does exist, and it is not easily bridged by the novice instructor. The literature, directed from educational technologist to educational technologist, is filled with specialized jargon and abstract presentations which shut out the uninitiated. Yet the uninitiated are the ones who must do what that little arrow represents: transpose job task data into behavioral objectives, and then into training texts. Military managers and educational technologists must provide instructors with the means to bridge the chasm.

At the Coast Guard Institute our staff comprises scarce few educational technologists and many novice instructors. The instructor supervisors are officers who have subject matter knowledge, but no training in curriculum development or text writing. Our staff performs dual functions of competitive exam development and correspondence course development. Each instructor is responsible for developing competitive exams for pay grades E-4 through E-6 semiannually and E-7 through E-9 annually. The cycle for the exams is rigid. Consequently, training course development must be worked in around the exam development schedule. Downstream bottlenecks further complicate the situation by limiting training course production to about ten courses per year. These two factors, dual function and downstream bottlenecks, mean that only about one third of our instructors are working on course development actively. Put another way, the supervisors and editors are deeply involved in exam review and are unable to devote their full efforts to course development. They are also unable to devote extensive on-the-job time to in-house training because of other demands. Nor can instructors be spared to a full time off-the-job course of instruction in educational methods and theories. Without such a program, though, the instructors, who are all senior supervisory technicians, are forced to work at a disadvantage. Our task over the past year has been to devise a program which will allow our instructors to bridge the gap from job task analysis to behavioral objectives without losing their services or diverting their supervisors.

I should digress for a moment here to note that because the Institute is a correspondence course center, course development consists of curriculum design and text writing. There is no actual contact with students. Our instructor's job requires a level of perfection which is not required of a classroom instructor. Classroom instructors can quickly correct an oral presentation, even modify it in progress. But once a text is printed it will remain in use, unchanged, for a long period of time.

Early in the process of instructor training, we realized that it was a mistake to ignore the training and experience which the instructors had gained in their ratings. This is one of the main attributes which we seek in recruiting instructors. Yet for some time we disregarded it as we tried various instructor training programs. We concentrated on training for job-entry into the instructor field, disregarding the fact that virtually none of our instructors had previous experience in that field nor remained in it after their tour at the Institute. Because they have devoted the better part of twenty years progressing from job-entry in their ratings through the journeyman level to the supervisory level, they are very much committed to the rating. Interest in educational methods and theories is initially very low. Rating knowledge, on the other hand, is very high. We decided to make this well honed rating knowledge the focal point of our program. We would initially set aside educational concerns and concentrate on what the instructor already knew.

At this point we began experimenting with the "by the numbers" approach to curriculum development. The original intent of this approach was to allow the new instructor to work on his own on a very limited portion of curriculum development. Progress beyond that limited portion would not be allowed until supervisors reviewed the work at specified stages. This was primarily designed with the two-fold purpose of ensuring proper work and eliminating instructor frustration. Frustration was a major concern because instructors earlier had been allowed to proceed too far without supervision. Often the result was rejected work. Time and effort were lost, and initiative was dulled. Closer review would be to everyone's advantage, but the supervisors' time had already been spread very thin.

Prescribing a "by the numbers" process ensures that the supervisor can review a completed stage of development rather than poring through several partially completed stages. Each step is described in a Course Development Manual so that the instructor will have a reference book related to his own job. The manual, short on theory but long on procedures, helps him to realize when he is doing any part of his work out of sequence. The finished product will be a text outline which is so thorough that he will easily be able to write a lean, effective manuscript.

The starting point is the job task analysis. Our interest for this process is confined to three types of analyses: percentage of people performing the task, percentage of time spent on the task, and grade level of people performing the task. The Enlisted Rating Qualifications Manual provides guidance on which tasks are required for each grade level, and the instructor's experience fills in the gaps that may have been left in preparation of the job task questionnaire. Using the data on numbers performing and time spent on each task, the instructor must first select out those tasks which cannot be taught in a correspondence course. Tasks which require practice, manual dexterity, close supervision, or "feel" are dropped out. This review also eliminates those tasks which should not be taught in a correspondence course. The target here is the list of tasks for which on-the-job training and drills have been prescribed by the Commandant or the District Commanders.

Two other decisions must be made during this stage. A cut-off point must be selected to separate those tasks performed often enough to be taught and those performed too seldom to be worth teaching. The second decision also requires identifying those tasks in the "too seldom performed" category which, though seldom performed, are critical on the job. The amount of resultant text material is one of the main criteria for fixing the cut-off point. The services of a GS-9 Training Specialist are available during this phase to assist the instructor in deciding whether text can be developed on a task, and if so, how much text will result. When this step has been completed, the supervisor will review the list of tasks to be taught.

READY, TWO

Using the "paygrade performing" form of the job task analysis and the Enlisted Rating Qualifications Manual (Quals Manual), the instructor will assign each of the tasks to the proper paygrade, E-4 through E-7. The job task analysis will tell him which paygrade is performing the tasks and the Quals Manual will tell him which paygrade must be able to perform it. If the task is required at a lower grade level than that commonly performing it, the task will be taught at the lower level. It may also be taught at a lower level than indicated by either the job task analysis or Quals Manual if a more efficient progression of rating courses will result. Often we find that many of the things required at the E-6 level on the job can be better taught at the E-5 level. Because enlisted duty assignments in the Coast Guard are not rigidly controlled by paygrade, we are able to make this type decision without discomfort. Instructor experience and judgment are just as important at this stage as the job task analysis and the Quals Manual.

During this stage the instructor must often revise step one, adding or deleting material. Courses for four grade levels will be required. If a grade level becomes flooded with material to be taught, the text will be too cumbersome to be completed by the student. Adjustments are made by shifting some material to a lower level course or by deleting some of the tasks selected in step one. If too little material results for a grade level, additional tasks may be added to the original list. The supervisor will review this stage when it has been completed.

READY, THREE

The supervisor will specify which of the four grade level courses is to be developed first. The instructor then takes his list of tasks for that grade level and begins several steps of organizing and subdividing.

Tasks listed on the job task analysis are in the order dictated by the responses of participants in the job task survey. To be useful in writing text material, the tasks must be regrouped. The tasks are first grouped into broad categories, or topics, based on the similarity of the tasks. Because this is primarily a clerical chore and is very much within the realm of expertise of the instructor, there is no review required before the next stage is begun. When the instructor is satisfied, he moves on the next stage.

READY, FOUR

This stage is brief and, again, primarily clerical. All of the tasks within each topic are put into a logical order. The instructor has free choice in deciding on the order, arranging the tasks from simple to complex, in the order they are performed, or any other order which he feels is logical for the topic. He proceeds independently to the next stage.

READY, FIVE

Once again the instructor is operating within the realm of his experience and expertise. Each of the tasks is broken down into all of the steps necessary to complete the task. This subdivision continues until the list of steps, or Job Subtasks, cannot be broken down any further. The supervisor reenters the process and reviews the third, fourth, and fifth stages.

READY, SIX

For the first time the instructor is going to operate out of the realm of his expertise, though not necessarily out of the realm of his experience. In this sixth stage, he will convert his list of _Job_ Subtasks into a list of _Lesson_ Subtasks. So far he has been dealing only with what is required on the job, where there are boats, engines, tools, typewriters, radars, radios and other similar _things_. Now he must convert the list of job subtasks to lesson subtasks, because in the lesson the only _things_ the student will have are paper and pencil. Each lesson subtask must be a paper-and-pencil version of a job subtask. Generally, lesson subtasks will be statements of the decisions which are required to perform the job subtask. On the job the student may be required to sharpen tools; in the lesson he should be required to select the proper grinding tool, select the proper angle of the tool edge, or identify a properly sharpened tool in a display of two or more tools. On the job he may be required to repair a diesel engine; in the lesson he should be required to identify repair procedures, or tools, or symptoms of proper and improper operation. Every job task requires decision making. These decisions provide the material to draw on for lesson subtasks.

This step is based on the assumption that if the student is able to perform the lesson subtasks successfully, he will probably be able to perform the job subtasks successfully. Performing these lesson subtasks constitutes the primary difference between studying a correspondence course and reading a reference book. Just as the classroom teacher requires certain activities from his students, so the correspondence course instructor requires activities from his students. The Training Specialist works closely with the instructor on this stage. Because each Training Specialist works with about ten instructors, he is able to carry over many ideas from course to course. The supervisor reviews the work at the end of this stage.

READY, SEVEN

The instructor writes the end-of-course (criterion) test at this point. In the original plan this stage was not included. After several revisions to the process, however, its placement was "obvious." Successful completion of the course, a prerequisite to advancement, is achieved by obtaining a passing score on the end-of-course test. Because it is a prerequisite to advancement, the test must be highly job relevant. To be useful, it should also be a reasonably good predictor of the likelihood of on-the-job success.

It is fairly easy at this stage to construct such a test by selecting lesson subtasks to be carried into the end-of-course test. The test will contain 100 multiple choice items based on 100 or fewer of the lesson subtasks selected for the test outline. Because the work of ensuring that lesson subtasks are paper and pencil exercises has already been completed and reviewed, writing the actual test items is easy. As each segment of text is written later on, the instructor can design student activities and end-of-course test items for that segment of the text. By designing the student activities and end-of-course test items at the time the text is developed, we can have a greater expectation that the test item will be a useful sample of the student's understanding of the text. This stage is reviewed by the supervisor and by the staff Testing Psychologist to ensure that the outline represents a realistic sampling of course knowledge.

READY, EIGHT

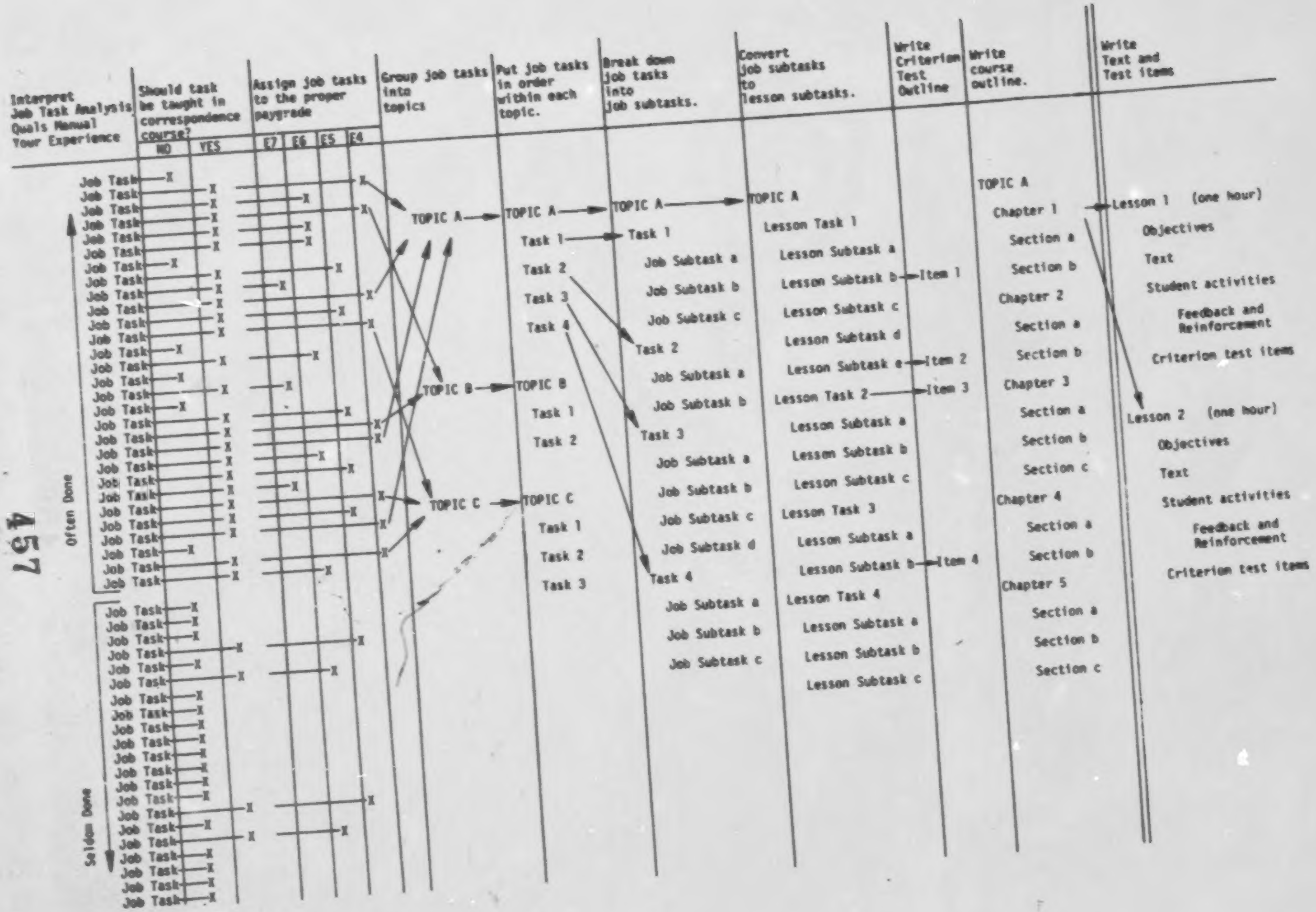
The instructor must now write an outline for the text manuscript. He studies his list of job and lesson subtasks and writes a chapter by chapter outline for each of the books to be included in the course. He will organize his outline so that he can write a unit consisting of text and student activities which the student will complete and score himself. Each unit is designed to be completed in one hour or less. When the supervisor approves the course outline, the instructor begins writing his text.

AT EASE

During the seventh and eighth stages we have been routinely receiving a surprise benefit. We had expected that the instructors would develop very good text outlines and that they would be conversant with a multitude of details for the text. We did not expect, however, that they would be discussing those details in terms of enabling behavior and terminal behavior. We had hoped to finesse behavioral objectives by developing our product "by the numbers." Much to our delight this approach had provided the new technician-turned-teacher with the basis for a clear understanding of the meaning and use of behavioral objectives. As he proceeded from step to step in the outline development process, he was working with concrete tasks from the job and for the lesson. Because the tasks were directly relevant both to his experience and to his project, the "light" quickly clicked on.

We had been unsuccessful in teaching the use of the systems approach based on behavioral objectives in advance of curriculum development. The instructor, on the other hand, by taking a retrospective view of his own work, comes to a rapid understanding with very little effort on anybody's part. Meanwhile, the developmental work for his first course has reached the manuscript stage with minimum lost time and virtually no wasted effort. Some small amount of time is expended by the supervisor and the Training Specialist in training the instructor at the various stages. The training is highly efficient, though, because it is narrowly focused and relevant to work in progress. Generalities and abstractions are avoided.

Supervisors have all reported that as each instructor goes into the development of second and third courses, there is less and less to criticize in the various review stages. This is a clear indication that the method works for the benefit of everyone--instructor, supervisor, and of course, the student who will be confronted with a responsibly prepared training course.


Interpret Job Task Analysis Quals Manual Your Experience
Should task be taught in correspondence course?
NO
YES
Assign job tasks to the proper paygrade
E7
E6
E5
E4
Group job tasks into topics
Put job tasks in order within each topic.
Break down job tasks into job subtasks.
Convert job subtasks to lesson subtasks.
Write Criterion Test Outline
Write course outline.
Write Text and Test items
Job Task
X
Often Done
Seldom Done
TOPIC A
TOPIC B
TOPIC C
Task 1
Task 2
Task 3
Task 4
Job Subtask a
Job Subtask b
Job Subtask c
Job Subtask d
Lesson Task 1
Lesson Task 2
Lesson Task 3
Lesson Task 4
Lesson Subtask a
Lesson Subtask b
Lesson Subtask c
Lesson Subtask d
Lesson Subtask e
Item 1
Item 2
Item 3
Item 4
Chapter 1
Chapter 2
Chapter 3
Chapter 4
Chapter 5
Section a
Section b
Section c
Lesson 1 (one hour)
Lesson 2 (one hour)
Objectives
Text
Student activities
Feedback and Reinforcement
Criterion test items

QUALITY OF LIFE IN THE U.S. AIR FORCE

T. Roger Manley, Air Force Institute of Technology
Robert A. Gregory, Directorate of Human Resources
Development, HQ USAF
Charles W. McNichols, Air Force Institute of Technology

In the Spring of 1975 the Chief of Staff of the U.S. Air Force, General David Jones, directed that a special study group be established to examine the Air Force's present "people" programs and to determine how well they will fit the Air Force of the future. Christened AFMIG (Air Force Management Improvement Group), the then Director of Personnel Plans and current Deputy Chief of Staff/Personnel, Lieutenant General Kenneth L. Tallman, was named study group director and given the broad charter of making "...a good service better."

The composition of the study group constituted something of a departure from the Pentagon's usual approach to studies in that Air Force members of various grades, job specialties, and backgrounds were drawn from the field to complement Air Staff personnel professionals. The authors were chosen to participate in the study group because of their academic, research and consulting backgrounds, and joined the effort after the initial groundwork had been accomplished and a general direction set. Along with two other Air Force members, they formed the research and analysis section of AFMIG, and were given the responsibility of functioning as resource persons for group members working on various tasks and for developing whatever questionnaires might be needed for the study.[1]

MGM THEORETICAL MODEL

In the initial flurry of activity, study group members identified a large number of issues to be investigated (the number ultimately reached 39). The issues covered such diverse subjects as considering changing life styles, the balance between permissiveness and individual responsibility, the need for leaders to retain decision-making powers versus the desire for autonomy on the part of Air Force members and their rights to appeal the decisions of supervisors, the realities of our current and future economy versus the needs of Air Force members for reasonable stability in their careers and equitable compensation, as well as a host of other controversial topics. These as well as the other issues which were surfaced reflected the basic premise adopted by General Tallman at the outset of the study; i.e., there has been substantial change in behavior, life styles, human relationships, attitudes toward the military and government, and the expectations of citizens, and these changes are having and will continue to have a significant impact upon the Air Force.

Perhaps reflecting our professional experience and educations, we soon began to search for some underlying theme or theoretical model which would provide a logical underpinning for the activities of the group. After some discussion we determined that what the study group was attempting to examine was the Quality of Air Force Life (QOAFL). Drawing upon the literature, and particularly the Environmental Protection Agency's 1973 QOL Symposium at Airlie House,[2] as well as upon previous behavioral science research within the Air Force, we identified nine factors which seemed to encompass the major thrusts of the study group. They were: (1) ECONOMIC STANDARD; (2) ECONOMIC SECURITY; (3) FREE TIME; (4) WORK; (5) LEADERSHIP/SUPERVISION; (6) EQUITY; (7) PERSONAL GROWTH; (8) PERSONAL STANDING; and (9) HEALTH. (See Figure 1 for definitions.)

In the survey on which this paper is based, a 150 question instrument was distributed Air Force-wide. The questionnaire was divided into nine sections, each related to a QOAFL factor, and contained 17 questions relating to personal and organizational variables. A Porter-Lawler approach was used for the rating of each of the nine factors, with respondents requested to rate each factor in terms of importance and degree satisfied.[3] Figure 2 shows how the factor FREE TIME was presented in the questionnaire. Questions related to each factor were placed immediately after the factor rating questions, and generally utilized five point Likert responses. As can be seen from Figure 3 by combining the importance dimension with the satisfaction dimension, we can examine each factor by using a 3 x 3 matrix. From the point of view of the policy-maker, any factor which fell along the diagonal (high-high, moderate-moderate, low-low) clearly represented no problem. The two cells which would be of greatest interest were, in our opinion, the high importance-high satisfaction and high importance-high dissatisfaction cells. The former would be significant from the viewpoint that all was well, and the latter would constitute clear cause for concern.

Further, we had expected _a priori_ that the high importance column, if we had selected relevant QOAFL factors, would contain most of the responses. As will be seen later, this indeed was the case. Since responses in the high importance-high satisfaction cell were clearly "good" and those in the high importance-high dissatisfaction cell clearly "bad," the question arose concerning the significance of the high importance-moderate satisfaction cell. We reasoned that responses falling in this cell were indicating some degree of dissonance. In our pretest of the questionnaire we questioned respondents who had responded in this manner on some of the factors, and our assumption appeared to be borne out -- although the degree of dissonance appeared to vary among respondents.

Figure 1

QOAFL FACTORS

ECONOMIC STANDARD: Satisfaction of basic human needs such as food, shelter, clothing; the ability to maintain an acceptable standard of living.

ECONOMIC SECURITY: Guaranteed employment; retirement benefits; insurance; protection for self and family.

FREE TIME: Amount, use, and scheduling of free time alone or in voluntary associations with others; variety of activities engaged in.

WORK: Doing work that is personally meaningful and important; pride in your work; job satisfaction; recognition for my efforts and my accomplishments on the job.

LEADERSHIP/SUPERVISION: Has my interests and that of the Air Force at heart; keeps me informed; approachable and helpful rather than critical; good knowledge of the job.

EQUITY: Equal opportunity in the Air Force; a fair chance at promotion; an even break in my job/assignment selections.

PERSONAL GROWTH: To be able to develop individual capacities; education/training; making full use of my abilities; the chance to further my potential.

PERSONAL STANDING: To be treated with respect; prestige; dignity; reputation; status.

HEALTH: Physical and metal well-being of self and dependents; having illnesses and ailments detected, diagnosed, treated and cured; quality and quantity of health care services provided.

In addition to treating the importance and satisfaction responses separately in our processing of the data, we also placed individual responses in the matrix form (Figure 3) and aggregated that data for the entire sample. Both forms of the data were used in our analysis.

Figure 2

QUESTIONNAIRE EXAMPLE: FREE TIME

SECTION III: FREE TIME

Numerous aspects of your life are centered around free time activities. Below is a definition of "free time" as we see it. We are examining the satisfaction of Air Force members with various free time aspects of their lives. Related topics are also included in this section.

FREE TIME: Amount, use, and scheduling of free time alone, or in voluntary associations with others; variety of activities engaged in.

What degree of importance do you attach to the above? (Select one of the seven points.)

A....B....C....D....E....F....G
Low Importance   Neutral   High Importance

To what degree are you satisfied with the FREE TIME aspects of your current life?

A....B....C....D....E....F....G
Highly Dissatisified   Neutral   Highly Satisfied

RESULTS

1. The Sample. Completed questionnaires were received from 10,996 Air Force members, ranging in grade from Airman Basic (E1) to Colonel (06). The sample was stratified by grade, and respondents randomly selected. The questionnaire was administered to respondents at their base of assignment, and anonymity was guaranteed. In no way could a response be traced back to an individual. In addition to the structured responses, approximately 8,000 respondents provided additional qualitative comments. Some other facts of interest concerning the sample are:

*Seventeen percent of the respondents were officers, 83 percent were enlisted.

*Eleven percent of the enlisted respondents had less than one year.

of service, as did five percent of the officer respondents. All year groups up through 30 years of service were represented, with less than one percent falling in the 30 years or more category.

*Eighty-two percent of the officers were married, as were sixty-eight percent of the enlisted respondents.

*Three percent of the officers sampled and fifteen percent of the enlisted were black. Other racial minority groups made up five percent of the officer sample and eleven percent of the enlisted sample.

*Women accounted for five percent of the officer sample and six percent of the enlisted sample.

*All educational levels were represented in both the officer and enlisted samples, including some individuals with less than a grammar school education and some individuals in both samples with a doctoral degree.

Figure 3

ANALYSIS MATRIX

	High Importance	Moderate Importance	Low Importance
Highly Satisfied			
Moderately Satisfied			
Highly Dissatisfied			

(Note: Ref. Figure 2, A,B = Highly Dissatisfied, Low Importance; C,D,E = Moderate Importance, Moderately Satisfied; and F,G = Highly Satisfied, High Importance.)

2. Quality of Air Force Life (QOAFL). Figure 4 presents each of the nine QOAFL factors in descending order of the percentage of respondents rating each factor as being of high importance. Also shown in the figure are the responses to the three different levels of satisfaction.

Figure 4

RATING OF QOAFL FACTORS
(N=10,996)

	High Importance	Hi Imp-Hi Sat.	Hi Imp-Mod Sat	Hi Imp-Hi Dissat
HEALTH	90%	38%	38%	14%
WORK	87	31	38	18
EQUITY	83	28	41	15
PERSONAL GROWTH	83	26	44	12
ECONOMIC STANDARD	81	18	49	13
ECONOMIC SECURITY	80	31	40	10
PERSONAL STANDING	79	29	39	11
LEADERSHIP/SUPERVISION	68	21	33	14
FREE TIME	64	22	31	12

(Note: Percentages are rounded-off to nearest whole number.)

In examining the data presented in Figure 4, a couple of points will probably attract the reader's attention. First, only three of the factors (HEALTH, WORK, and ECONOMIC SECURITY) were placed in the high importance-high satisfaction cell by thirty percent or more of the respondents. Since this is a new approach to examining QOL, the authors are not quite certain as to whether this is healthy or not. As will be discussed later, the formal analysis of the data is in the very early stages and we are not quite certain of what we will find. (As this article is being written, the data is less than one month old.) It is interesting to note, however, that the high placement of

HEALTH is consistent with Dalkey's finding in 1968 using the Delphi approach.[4] Also, it seems to us that we have a potentially valuable set of baseline indicators, providing a gross indication of the quality of life in the Air Force, and their primary value may well prove to be in the area of detecting change.

With the current interest in worker alienation and job enrichment/enlargement, it is also interesting to note that the factor providing the highest indication of dissatisfaction was that of WORK. Again, we do not know if 18 percent of the respondents indicating that their work is highly important to them but that they are highly dissatisfied with it is a good or a bad sign. Intuitively it would seem that having one out of every five members turned-off by their work, which they indicate is highly important to them, indicates that some form of job enrichment is called for. However, we cannot justify that conclusion at this point; in a relative, or perhaps even absolute sense this proportion might be quite acceptable.

Another perspective on these data which is of interest is the comparison of officer and airman responses. This is presented in Figure 5.

Figure 5

RATING OF QOAFL FACTORS: OFFICERS vs. AIRMEN
(N=1957 / N=9039)

	High Importance	Hi Imp-Hi Sat	Hi Imp-Mod Sat	Hi Imp-Hi Dissat
HEALTH	93%/ 89%	36%/ 38%	43%/ 38%	14%/ 14%
WORK	94 / 86	38 / 29	40 / 37	16 / 19
EQUITY	90 / 82	36 / 26	40 / 41	14 / 15
PERSONAL GROWTH	91 / 81	35 / 24	44 / 45	12 / 12
ECONOMIC STANDARD	80 / 80	37 / 15	39 / 52	4 / 15
ECONOMIC SECURITY	81 / 80	31 / 31	40 / 40	10 / 10
PERSONAL STANDING	76 / 80	33 / 27	33 / 40	15 / 13
LEADERSHIP/SUPERVISION	83 / 65	29 / 20	39 / 32	15 / 13
FREE TIME	67 / 64	30 / 22	34 / 30	11 / 12

(Note: Percentages are rounded-off to nearest whole number.)

Some points of interest about the data presented in Figure 5 are: (1) a higher proportion of officers are in the high importance-high satisfaction cells for seven of the nine factors; (2) officer and airman responses are almost identical for the factor ECONOMIC SECURITY, however, officers seem much more satisfied with ECONOMIC STANDARD; (3) although airmen consider PERSONAL STANDING more important than officers, they are much less satisfied with this aspect of their Air Force lives; (4) although officers consider LEADERSHIP/SUPERVISION more important and are more satisfied with it than airmen, a larger proportion of officers than airmen are at the same time dissatisfied with it; (5) officers appeared to be substantially more satisfied with EQUITY and PERSONAL GROWTH than did airmen; and (6) while a higher percentage of officers considered WORK highly important, airmen were more dissatisfied with this aspect of their Air Force lives. The ratings assigned FREE TIME were also of interest, although the interest stems from the relatively low importance ratings it received from both officers and airmen. On the basis of material encountered in the literature, we had expected FREE TIME to be competitive in importance with the other factors.

3. Job Satisfaction as a Criterion Variable. Included in the survey was a form of Hoppock's general job satisfaction blank, modified slightly to be more appropriate to Air Force respondents.[5] This measure of job satisfaction contains four questions, each having four possible responses. The most negative response is assigned a weight of "1", and the most positive response a weight of "7". The four responses are summed to provide a general measure of job satisfaction. Respondents ranged from the lowest possible score (0.8% had scores of 4), to the highest possible score (0.7% had scores of 28).

Two of the authors had used this measure of job satisfaction in various research and consulting efforts within the Air Force since 1972. The mean job satisfaction score accumulated through these studies (N=1331) was 17.85. The mean job satisfaction for the respondents to this survey was 17.69. This figure was somewhat of a surprise to us because the previous data were drawn primarily from officer personnel and civilian research and development scientists and engineers. We had anticipated that the inclusion of a large number of enlisted personnel would result in a significant drop in the mean job satisfaction score. Although airmen did in fact have a lower mean job satisfaction score than officers (17.48 vs. 18.71), the drop was not of the magnitude expected.

ANALYSIS

Since there are many questions in the survey providing detailed information related to each of the nine QOAFL factors, data is available to support a variety of specific analyses. As mentioned earlier the data discussed in this paper are less than a month old and the analysis which has been completed to date is of a preliminary nature.

1. Correlation Analysis. As a preliminary interpretive step, we have examined the correlation structure of the nine QOAFL factors and their relationship to Hoppock's job satisfaction measure as well as to respondent indication of career intent. Career intent was presented to respondents by five possible responses ranging from definite plans to make the Air Force a career to definite plans to separate.

Job satisfaction and career intent had a correlation coefficient of .35, which is somewhat lower than might have been expected. Figure #6 summarizes the three highest and three lowest correlations between the QOAFL factors and job satisfaction and career intent. For purposes of this analysis correlations were derived using the "degree of satisfaction" associated with each of the nine factors.

FIGURE 6

FACTORS MOST AND LEAST CORRELATED WITH JOB SATISFACTION AND CAREER INTENT

	Job Satisfaction		Career Intent	
	factor	correlation	factor	correlation
Highest Correlations	WORK	.72	PERSONAL STANDING	.36
	PERSONAL GROWTH	.43	WORK	.33
	PERSONAL STANDING	.40	PERSONAL GROWTH	.33
Lowest Correlations	ECONOMIC STANDARD	.20	FREE TIME	.16
	ECONOMIC SECURITY	.19	ECONOMIC SECURITY	.14
	HEALTH	.17	HEALTH	.08

Canonical correlation analysis was used to find the linear combination of the nine QOAFL factors, and the linear combination of job satisfaction and career intent which would be most highly correlated. By using this technique relative weights were obtained for the nine factors, yielding an index with the highest possible correlation with a composite job satisfaction/career intent measure. The actual canonical correlation obtained in the analysis was .74. Figure #7 presents the coefficients obtained from this analysis.

FIGURE 7

CANONICAL CORRELATION ANALYSIS

Factor	Coefficient
WORK	.437
PERSONAL GROWTH	.081
PERSONAL STANDING	.070
LEADERSHIP/SUPERVISION	.035
HEALTH	-.025
EQUITY	.016
ECONOMIC SECURITY	.007
FREE TIME	.004
ECONOMIC STANDARD	.004
Career Intent	.138
Job Satisfaction	.185

Note in Figure 7 that WORK, PERSONAL GROWTH, and PERSONAL STANDING are the factors with the highest weights, and that WORK is weighted more than five times as heavily as the second most significant factor. Also, in the composite measure, job satisfaction is weighted somewhat more heavily than career intent.

2. Automatic Interaction Detection (AID). An analysis of respondent job satisfaction scores is nearing completion and will be reported in detail at a later time. In our job satisfaction analysis, demographic factors and the structured questions related to the WORK and LEADERSHIP/SUPERVISION QOAFL factors were examined using the Automatic Interaction Detection (AID) algorithm.[6] We are very enthusiatic over AID's apparent ability to provide a simple description of high and low job satisfaction groups in the Air Force. Our analysis to date has shown that the level of job satisfaction is strongly associated with individual responses on the following three parameters:

a. Degree of challenge offered by the respondent's current job (whether the respondent desires challenge or not).

b. The degree to which the respondent perceives his or her current job as preparing him/her for positions of greater responsibility.

c. The degree of freedom given the respondent in accomplishing his/her current job.

Job satisfaction increases with increasing scores on any of the three parameters, and this relationship holds for both officer and enlisted personnel. Furthermore, an intuitively appealing and statistically significant multiple regression equation has been constructed relating job satisfaction to a linear combination of the three parameters.

DISCUSSION

During the past several years we have observed a growing and intense interest in the concept of the quality of working life. While the interest is some quarters appears to stem from the much publicized concern over growing alienation of workers from their jobs and the belief that the way to increase productivity is to decrease this alienation, many more authorities seem to recognize it as an idea whose time has come. Further, as the editors of the EPA's QOL Symposium proceedings point out, "to...promote the general welfare" is one of the statement of purposes contained in the preamble to the Constitution, and as such is a constitutional basis for recent (and growing) government QOL initiatives.

Although a number of writers have provided their definitions of QOL, no single definition has won widespread acceptance. QOL is viewed differently by different groups, e.g., youth-adults-senior citizens, men-women, whites-blacks-Mexican Americans-American Indians, poor-middle class-wealthy, etc. In years past QOL was normally restricted to issues of pay, length of the work day, and working conditions. However, today the term has much broader connotations. As members of the 1971 White House Conference on Youth observed in their report:

> In our nation of rising and unprecedented affluence, social discontent runs deep and as yet unmeasured. The vague areas of dissatisfaction form the heart of the problem known as the "quality of life" question.

The Center for Organizational Studies at UCLA, working under contract to the U.S. Department of Labor, developed an extensive annotated QOL bibliography of the literature from 1957-1972.[7] Their major criterion for inclusion of a work was that it deal with one or more of the following QOL parameters: (1) Alienation; (2) Health and Safety; (3) Economic Security; (4) Self-Esteem; (5) Self Acutalization -- Learning and Growth, and Using Existing Competence; (6) Work Environment -- Physical and Social; (7) Control and Influence; (8) Organizational Enclosure; (9) Career Aspirations; Extra-Work Activities; and (10) Home and Family. As can be appreciated from this list (and the voluminous size of the bibliography) QOL touches upon virtually every significant aspect of an individual's life.

In our research we were interested in a specific subset of the American population: the men and women on active military duty in the

U.S. Air Force, and appropriated fund civilian employees of the Air Force. In particular, we were interested in those facets of Air Force life about which the Air Force leadership could do something. In short, we wanted a gross measure of the QOAFL, which would identify potential problem areas and which, when replicated, over time, would be sensitive to changing trends.

As to our conceptualization of QOAFL, we tend to agree with those researchers who view QOL as a "sense of well-being". Recognizing full well that we have ventured into highly subjective territory, we nevertheless believe that we have gained some insights which are and will be useful to Air Force policy leaders. One example of such insight deals with NCO's (E4-E9). In examining responses to the factor PERSONAL STANDING we observed that the more senior a respondent, the more important and the more satisfied the individual appeared to be with this aspect of his or her life. While this same pattern held for both officers and enlisted, the spread between enlisted grades was much more pronounced than for the officer ranks. In examining responses to questions which supported this factor, we discovered some interesting findings.

*The lowest ranking NCO, E-4, was more apt to respond similarly to lower ranking airmen (E1-E3). In those cases where the E-4's responses did not resemble lower ranking responses, they seemed to stand-off by themselves rather than relate to E-5 responses.

*Almost twice as many NCO's as officers expressed a belief that the grade they held was too low for the work they were doing (19% vs. 34%).

*Of all the grade groupings, senior NCO's (E7-E9) had the highest proporition of respondents in the high importance-highly satisfied cell for WORK (51%), and yet 50 percent reported that they are usually given jobs with less responsibility than they should have (a view which was not shared by respondents of other grades). Further, when asked whether senior NCO's were primarily supervisors or technicians, a higher proportion of senior NCO's considered themselves technicians than did respondents of other grades. Also, senior NCO's felt most strongly that the prestige and privileges of NCO's has been declining over the past several years.

We checked these findings out with various groups of NCO's, both at AFMIG sponsored workshops and by visiting NCO schools, and seemed to strike a wellspring of emotions. There was no doubt in our minds after the interactions between AFMIG members and the several hundred NCO's contacted, that feelings ran deep; NCO's in the Air Force are convinced that their prestige has declined and the privileges of their rank eroded. In examining the data for some indication of why these perceptions were so strongly held, the picture became more confused. We found, for example, that at least two out of three E1-E3 respondents stated that they respect most of the senior NCO's they know. Officers felt even more strongly on the subject; almost 90 percent of them reported respecting most of the senior NCO's they know. Furthermore, junior airmen and officers agreed with the senior NCO's; they too expressed the belief that the prestige and the privileges of the senior noncoms has declined.

As a direct result of these findings and the work of AFMIG members probing these issues, a number of initiatives have been taken as well as others being evaluated for possible implementation. The Chief of Staff of the Air Force has already approved a plan aimed at the revitalization of leadership training in the Air Force, and one of the prime areas of focus will be on the newly promoted E-4. A number of other initiatives, some still being formulated, are being given serious consideration for the NCO's. In short, we identified an area in which the sense of well-being of a subset of the Air Force population seemed less than it could be. Only time will tell whether or not the actions currently being taken or planned will succeed. The point is, however, a source of dissatisfaction was identified and its existence brought to the attention of Air Force leadership. Furthermore, on the basis of these findings Air Force leadership has chosen to take positive action.

On the basis of our early analytical efforts, the modified form of the Hoppock job satisfaction measure appears to possess significant discriminant power. As noted earlier we have used it as well as indication of career intent as criterion variables. At this point we have been able to identify three QOAFL factors and three variables which seem to have the greatest impact upon both job satisfaction and career intent. They are:

*WORK

*PERSONAL GROWTH

*PERSONAL STANDING

*Whether or not their present job is perceived as challenging

*Whether or not they perceive their present job as preparing them to assume increased responsibility.

*Whether or not they are given the freedom they need to do their job properly

On the basis of these initial findings it appears to us that the data will support the notion of the centrality of work to Air Force members. The work itself appears to be far more important to Air Force members than such considerations as Base Exchanges, golf courses, etc. Therefore, the Air Force may be well advised to undertake a systematic analysis of the design of jobs and explore the possibility of introducing job enrichment on a wide scale basis.

SUMMARY

This paper presents a preliminary examination of data recently gathered from an Air Force-wide survey designed to investigate the Quality of Air Force Life (QOAFL). The survey was undertaken by the Air Force Management **Improvement** Group (AFMIG) which was chartered by the Chief of Staff of the Air Force to take a far ranging look at ways of "making a good service better". In addition to the survey of active duty military members, AFMIG also conducted QOL surveys of all Air Force Base Commanders (N=115), and appropriated fund civilian Air Force employees (N=16,000+). At the present time a fourth study, examining the QOL of Air Force spouses, is being distributed in the field. The latter studies will be discussed in future reports.

In our attempt to capture an indication of the QOAFL, we presented nine factors which the 11,000 respondents evaluated both in terms of importance and degree of satisfaction. The factors (in order of importance) were: HEALTH, WORK, EQUITY, PERSONAL GROWTH, ECONOMIC STANDARD, ECONOMIC SECURITY, PERSONAL STANDING, LEADERSHIP/SUPERVISION, and FREE TIME. In using both the satisfaction and importance dimensions to rate the factors, three of the nine were rated high importance-highly satisfied by 30 percent or more of the respondents. They were: HEALTH, WORK, and ECONOMIC SECURITY. Of interest was the finding that while the factor WORK was second in importance only to HEALTH and had the second largest proportion in the high importance-highly satisfied category, it also had the largest proportion of respondents (18%) in the high importance-highly dissatisfied category.

Preliminary analysis of the data indicates that a central issue of concern for those individuals surveyed was the work itself. On the basis of what we have found to date, a systematic analysis of the design of jobs and possible implementation of job enrichment appears warranted.

Since this was the first time an attempt has been made to examine the QOAFL, the primary value of this effort may well be that it will establish a series of baseline measurements which, when replicated in the future, will provide indications of change in the QOL of Air Force members.

NOTES

1. The authors wish to acknowledge the contributions of the other two members of AFMIG's research and analysis section, Col (Dr) Joseph E. Zuro and Capt Paul Winkler.

2. The Quality of Life Concept: A Potential Tool for Decision Makers, U.S. Environmental Protection Agency, 1973.

3. Lyman W. Porter and Edward E. Lawler III, Managerial Attitudes and Performance (Homewood, Illinois: Richard D. Irwin Inc., 1968).

4. N. C. Dalkey, "Quality of Life," unpublished RAND paper, 1968. (A condensed version of the paper appears in reference 2., pp. II-192 - II-201.)

5. R. Hoppock, Job Satisfaction (New York: Harper and Brothers, 1935).

6. John A. Sonquist and James N. Morgan, "The Detection of Interaction Effects", Monograph No. 35, Institute for Social Research, The University of Michigan, 1964.

ASSIGNMENT LOCATION PREFERENCES OF AIRMEN

By

Joe T. Hazel
William J. Stacy
Lloyd D. Burtch
Occupational and Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland AFB, Texas 78236

I. INTRODUCTION

In the study of job satisfaction and career decisions, the military services encounter an unusual situation. Due to the large number of facilities, dispersion of locations, turnover and training, the military probably require more frequent reassignment of personnel than found in the civilian sector. Consequently, attitude towards assignment location is a dimension of job satisfaction research which merits special consideration.

There is consistent evidence substantiating the importance of assignment locations on job satisfaction, both from periodic surveys and other research concerned with the identification of dimensions of job satisfaction (Tuttle and Hazel, 1974; Tuttle, Gould, and Hazel, 1975). In addition, the importance of this dimension seems well recognized by the Air Force, and operational procedures have been established to fulfill assignment requirements, yet still give individual choices of assignments. While the desirability of overseas locations has been broadly defined and tour lengths established accordingly, fewer standards are available concerning the desirability of Continental United States (CONUS) bases or locations. Therefore, the purpose of this investigation was to examine the CONUS location assignment preferences of airmen and identify certain variables which may enter into the determination of assignment desirability.

In a recent study (Tuttle, Brockhaus & Hazel, 1974), a method was developed and tested whereby airmen could be surveyed systematically to determine which CONUS bases or locations they considered desirable and undesirable. In addition, the procedures permitted examination of certain geographical-environmental variables with potential influence on these preferences. Results of this study supported the feasibility of the method, however since the procedure was field-tested on a sample of basic airmen with limited time in service, further research on a sample of more experienced airmen seemed warranted. The present study is concerned with application of a modified version of this methodology to a large representative sample of enlisted personnel.

II. METHOD

The revised Airmen Location Assignment Preference Survey used for data collection consisted of six sections. Parts 1 through 4 contained a variety of background, attitudinal and preference items. Part 5 of the survey booklets contained 150 unnamed Air Force CONUS locations described in terms of 22 geographical or environmental characteristics. Part 6 consisted of the same 150 locations identified by base or station name and location. The scale used to obtain both location description and base name preference ratings consisted of 18-points ranging from extremely undesirable (-9) to extremely desirable (+9), omitting a zero point. Since data from these two sections were of primary concern in this study, particularly base name ratings, a few of the more prominent features concerning formats and rating scales warrant additional explanation.

Figure 1 partially illustrates the format used for collecting the Base Name Ratings from Part 6 of the survey. The format for presenting the 22 geographical-environmental variable descriptions for each of the 150 locations and method for obtaining preference ratings for these descriptions is shown in Figure 2. The 22 variables were selected because they were considered relevant to assignment preferences and represent actual descriptive information concerning each location, although other relevant variables may be identifiable.

Completed surveys were received from 4,670 airmen. Analyses were performed to investigate the representativeness of the response rates across different groups (i.e., grade, race, sex, MAJCOM). Findings indicated that the data appeared representative and did not contain any obvious sampling biases.

PREFERENCES FOR AIR FORCE LOCATIONS

You are now asked to give your opinions concerning assignment to the 150 locations, when each location name, city, and state is provided. The number sequence of the bases in this section does not correspond to the number sequence of Part V. Please indicate how desirable or undesirable an assignment would be for you to each location. Your rating may range from –9 meaning extremely undesirable to +9 meaning extremely desirable.

Be sure to write in the sign, either (–) or (+) and a number from 1 to 9 for each location. For example: If you decide that Aiken AFB, Aiken, SC was very desirable to you, then write a +7 at the end of the row for that location under the heading "RATING."

No.	Name and Location	Rating Rate each location from –9 to +9
1	Aiken AFS, Aiken, SC	
2	Air Force Academy, Colorado Spgs, CO	
3	Air Force Actg & Fin Ctr, Denver, CO	
4	Atlantic Cy Apt, Atlantic Cy, NJ	
5	Altus AFB, Altus, OK	
6	Andrews AFB, Camp Springs, MD	
7	Antigo AFB, Antigo, WI	

No.	Name and Location	Rating Rate each location from –9 to +9
144	Webb AFB, Big Spring, TX	
145	Westover AFB, Chicopee Falls, MA	
146	Whitman AFB, Sedalia, MO	
147	Wilder AFS, Wilder, ID	
148	Williams AFB, Chandler, AZ	
149	Wright-Patterson AFB, Dayton, OH	
150	Wurtsmith AFB, Ascoda, MI	

Figure 1. Partial Illustration – Part VI.

PREFERENCES FOR ENVIRONMENTAL/GEOGRAPHIC LOCATION DESCRIPTIONS

On the following pages you will find descriptions of 150 CONUS Air Force locations. Each location is unnamed but is described in terms of various environmental or geographical characteristics.

For each location read the characteristics and then assign a rating to indicate how desirable or undesirable an assignment would be for you to that location.

Your rating may range from –9 meaning extremely undesirable to +9 meaning extremely desirable. Be sure to write in the sign, either (–) or (+) and a number from 1 to 9 for each location.

Below is an example which indicates how you might read the information about a location.

EXAMPLE

LOCATION DESCRIPTIONS																							
No.	Base Military Population	Community Population Near Base	Miles to City of 200,000 or more Population	Cost of Living Relative to National Average	Educational Facilities (available within one hour driving time)			Distance in miles to:				Summer Weather					Winter Weather						Rating Rate each location from –9 to +9
					2-yr coll	4-yr coll	Grad/Prof	Mountains	Desert	Ocean	Lakes/River	Avg Jul Temp Hi	Avg Jul Temp Lo	Avg % Humidity	No. Days Rain-Fall Over .1"	Avg % Daily Sun-Shine	Avg Jan Temp Hi	Avg Jan Temp Lo	Avg % Humidity	No. Days Rain-Fall Over .1"	Avg % Daily Sun-Shine	Annual Snow-Fall (inches)	
1.	10,000	3,000,000	0	Slightly Below	Yes	Yes	Yes	150	1,500	1,000	10	80	60	75	18	65	43	29	65	22	34	111	+7
2.	500	5,000	40	Much Below	No	No	No	10	10	500	158	60	55	35	9	89	49	45	33	6	78	0	
3.	3,000	2,000	10	Average	No	Yes	Yes	5	110	25	5	76	62	67	10	78	55	39	40	36	67	2	

Figure 2. Partial Illustration – Part V.

III. RESULTS AND DISCUSSION

Table 1, presents the total sample means and standard deviations of base name ratings for the ten most preferred locations and the ten least preferred locations. Examination of the entire set of means revealed certain trends regarding desirable and undesirable assignment locations. Examination of all 150 location mean ratings also showed that there were many more locations with negative means than with positive means (i.e., 27 positives, 123 negatives).

Three plots on CONUS maps of the 50 most, intermediate, and least preferred locations are given in Figure 3. Inspection of the geographical areas in these plots indicated that the most preferred locations tended to be concentrated in the southeast (Florida particularly), California and Colorado. The 50 least preferred locations were in "northern tier" states, Texas, and several midwestern states. The 50 intermediate locations tended to be more widely scattered than the 50 most and least preferred locations. When the three plots were compared, there were relatively few intermediate locations which overlapped the most or least preferred area concentrations with the possible exception of the Texas-Arizona area (i.e., Texas had a considerable number of both intermediate and least preferred locations.

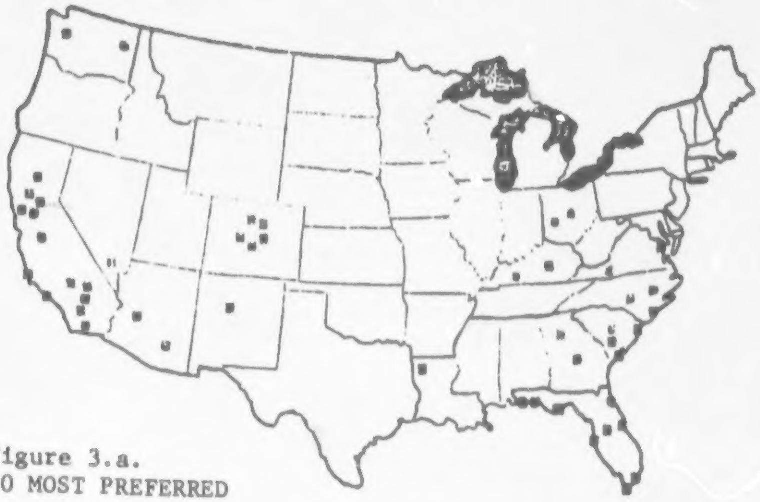
Preference Variability

A question of interest was the extent to which base name ratings varied as a function of preference level or desirability. To determine the relationship between levels of variability and preference, the means and standard deviations of the 150 locations were correlated ($r = .64$). This correlation indicated that the more desirable locations tended to have the larger rating variances and the less desirable locations the smaller variances.

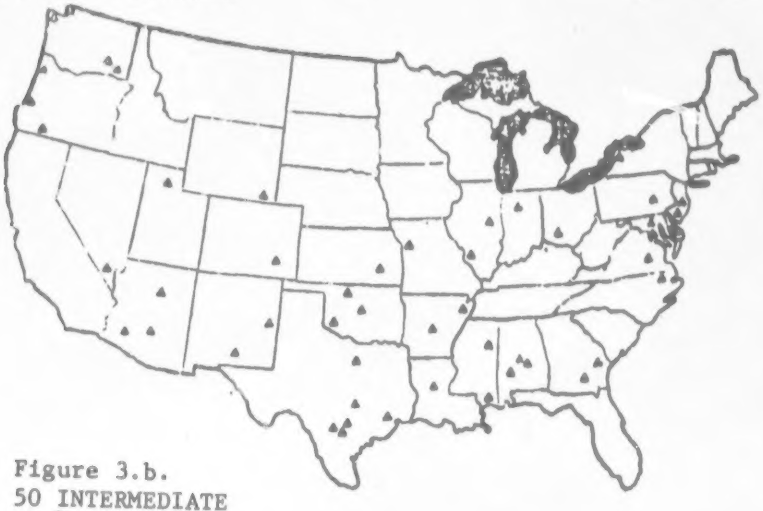
To examine more closely certain state or geographical area patterns involved in this variability and mean preference relationship, three CONUS plots were made of the 50 locations with the largest, medial and smallest standard deviations (see Figure 4). These plots were then compared to the mean plots in Figure 3 reflecting the following trends. The three states (Florida, California, and Colorado) with large concentrations of most preferred means had many locations with correspondingly large standard deviations. Numerous northern tier and midwestern area locations had

Table 1. Ten Most and Ten Least Preferred Locations For Total Sample

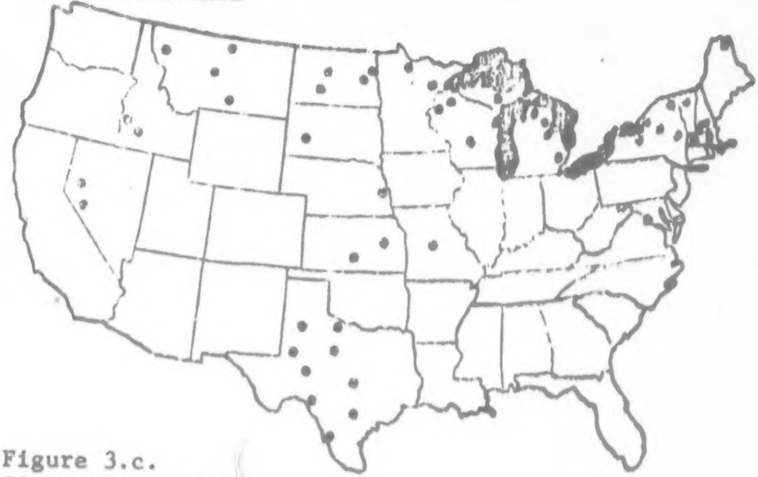
Rank/Mns	Base Name	M	SD	Rank/SD
1	Patrick AFB, Cocoa Beach, Fla.	+2.33	5.77	48
2	MacDill AFB, Tampa, Fla.	+2.29	5.73	54
3	Eglin AFB, Valpariso, Fla.	+2.08	5.77	47
4	McCoy AFB, Orlando, Fla.	+2.07	5.79	42
5	Homestead AFB, Homestead, Fla.	+2.03	5.73	53
6	Tyndall AFB, Panama City, Fla.	+1.87	5.81	39
7	Lowry AFB, Denver, Col.	+1.85	5.89	28
8	Eglin AF Aux 9 Fld, Valpariso, Fla.	+1.83	5.81	38
9	Jacksonville AFS, Jacksonville, Fla.	+1.73	5.75	50
10	Ent AFB, Colorado Springs, Col.	+1.70	5.73	55
.				
.				
.				
141	Kincheloe AFB, Kinross, Mich.	-3.40	5.19	135
142	Sault St Marie AFS, Sault St Marie, Mich.	-3.56	5.24	130
143	Havre AFS, Havre, Mont.	-3.60	5.12	140
144	K.I. Sawyer AFB, Gwinn, Mich.	-3.60	5.40	112
145	Ellsworth AFB, Rapid City, S.D.	-4.05	5.30	123
146	Loring AFB, Limestone, Maine	-4.08	5.28	127
147	Finley AFS, Finley AFS, Finley, N.D.	-4.82	4.75	150
148	Grand Forks AFB, Grand Forks, N.D.	-4.87	5.03	143
149	Bismark AFS, Bismark, N.D.	-4.96	4.88	147
150	Minot AFB, Minot, N.D.	-5.60	4.92	146



■ Figure 3.a.
50 MOST PREFERRED



▲ Figure 3.b.
50 INTERMEDIATE



● Figure 3.c.
50 LEAST PREFERRED

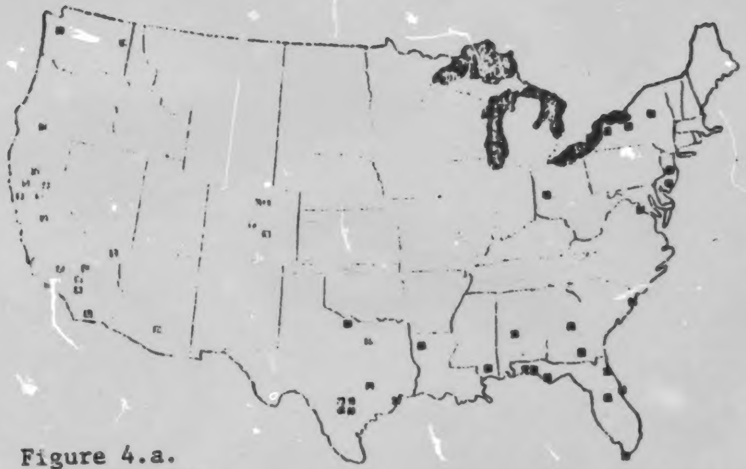
Figure 3. Plot of Fifty Most, Intermediate and Least Preferred Locations Based on Mean Base Nme Ratings

small standard deviations. An overall comparison of the three plots in the two figures revealed certain shifts or exceptions to the preceding trend which helped to explain why the relationship between mean preference and variability was not higher. Many of the Texas and northeastern bases which were in the lowest or intermediate mean preference categories had either the largest or medial standard deviations.

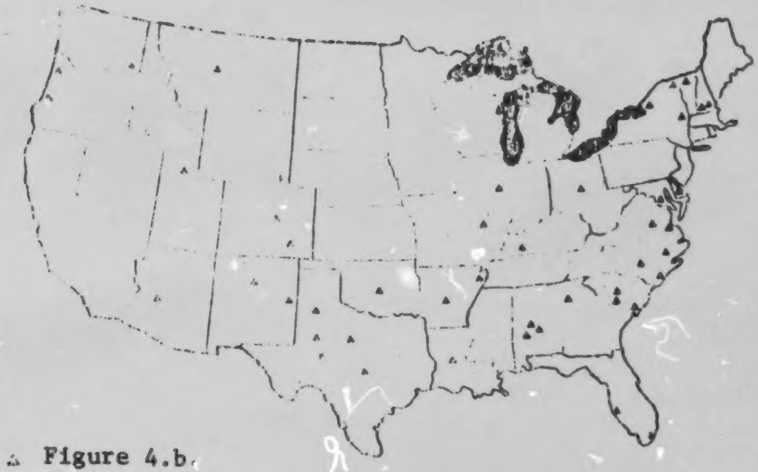
Comparison of Base Name Preferences for Various Sub-groups

Correlation analyses were used to examine the correspondence of order of mean base preferences for various subgroups. Subgroups were identified using factors such as career status, marital status, current location of assignment and job satisfaction.

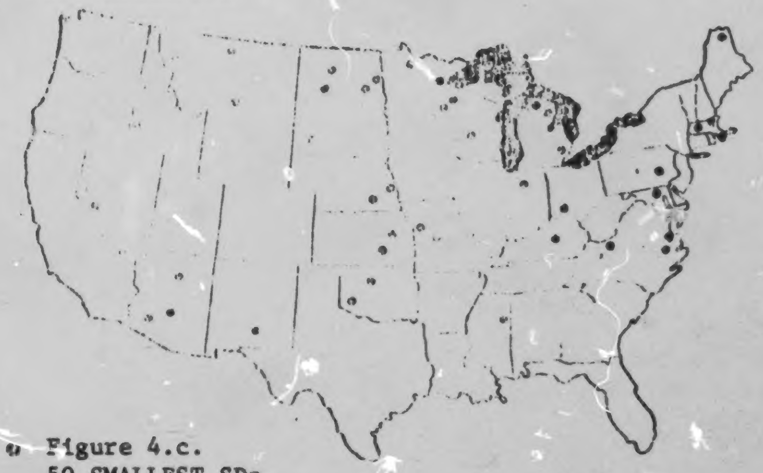
The correlations between the two groups assigned to either the 25 most or 25 least preferred locations was quite high (r = .98). Context of current assignment did not seem to influence the relative order of preferences for the 150 locations. The relationship between order of preferences for satisfied and dissatisfied airmen was of a similar magnitude (r = .97). For the married versus single and first-termer versus careerists however, the correlations were lower (r = .84 and r = .82 respectively). An examination of the rank orders of the sets of 150 means for these two groupings was made to determine if there were shifts in preferences for certain locations which accounted for the lower correlations. For first-termers versus careerists, a comparison of locations revealed the lower correlation was due to differences in preference for Texas bases. For first-termers most Texas bases were at the bottom of the preference listing (i.e., 12 Texas bases were among the 25 least preferred locations). For careerists, the "northern tier" bases were ranked lowest whereas some Texas bases were more towards the intermediate range and several locations (e.g., Carswell, Bergstrom, Ellington) among the 50 most preferred. For single versus married airmen, similar shifts in preferences were found. This similarity is understandable since the correlations between sets of preference means for first-termers and single airmen was .99. Another factor which may contribute to the "Texas shift" for first-term airmen is the number of large training facilities (e.g., Lackland and Sheppard) in this state, and the recent training experiences associated with such locations.



■ Figure 4.a.
50 LARGEST SDs



△ Figure 4.b.
50 MEDIAL SDs



● Figure 4.c.
50 SMALLEST SDs

Figure 4. Plot of Fifty Locations With Largest, Medial, and Smallest Variability In Base Home Ratings

Home and Family Influences on Preferences

A traditional or often mentioned explanation offered to account for assignment location preferences is the influence of "home and family ties" (e.g., place of birth, point of entry into service, and residence of relatives). For this reason, two variables concerning state home of record and residence of father, mother or spouse were investigated. In addition, a variable titled "state prefer to settle in after separation or retirement from the Air Force" was included in order to examine location preferences more directly. For all three variables, the cases selected for investigation were limited to airmen whose home of record, residence of father, mother or spouse, or "settle" statement was in one of nine least desirable "northern-most" states (i.e., North and South Dakota, Maine, Montana, Michigan, Minnesota, Wisconsin, Massachusetts, and Idaho).

Base name rating means for the 150 locations across all cases in each of the three groupings were computed. These three sets of means were then correlated with each other and with the set of means for the total sample.

The values of the correlation coefficients indicated that all three groups rated the 150 locations differently, when compared to the total sample ratings. The group for the variable titled "state prefer to settle in" reflected the largest shift in order of preferences (i.e., lowest relation with total sample, (r = 21), followed by home of record (r = .63), then residence of father, mother, or spouse (r = .84). These findings, when considered in conjunction with plots of high to low preferences reported subsequently, suggested that these two variables ("state prefer to settle" and "home of record") may be useful to identify airmen who may be satisfied with assignments to locations considered more undesirable by the total sample.

To examine trends in assignment location preferences for the two groups showing the lowest relation with the total sample, plots on CONUS maps were made for the fifty most, intermediate, and least preferred locations for each group. Plots of these preferences for the group of airmen who preferred to settle in the nine least desirable states are given in Figure 5. Plots for the group with home of record in these nine states are given in Figure 6. The group whose father, mother, or spouse resided in the nine least preferred states was similar to the home of record group (r = .94).

As shown in Figure 5, for the group of airmen who stated a preference to settle in the nine least preferred states, certain definite trends were observed. The 50 most preferred locations based on base name ratings tended to be concentrated in the Great Lakes states, in the far northwest, Colorado, and Florida. The 50 intermediate locations had a noticeable concentration in California and the southwest with some locations in North Dakota and the mid-central states. The 50 least preferred locations were concentrated in Texas and the south and southeast. When contrasted to the plots for the total sample, the most notable difference for airmen preferring to settle in the northern tier states is their low desirability for southern bases.

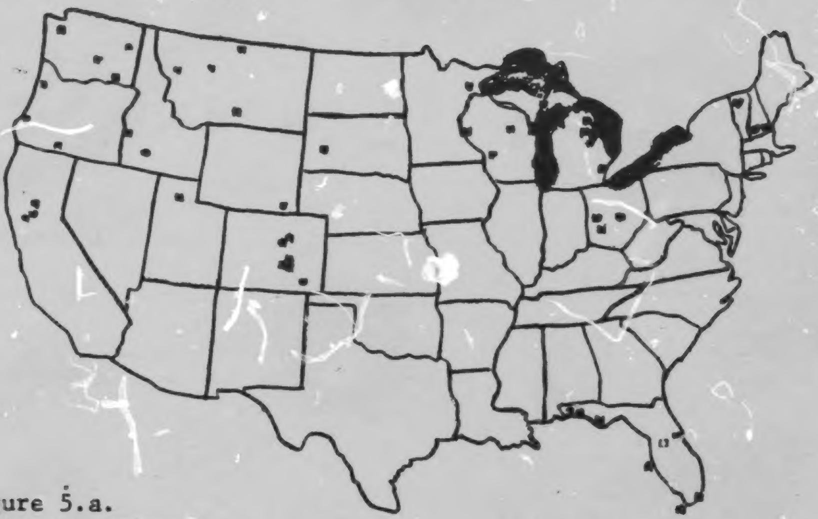
For the group with homes of record in the nine least preferred northern tier states, both similarities and differences in location assignment preferences were detectable, when compared to the total sample and to the group who preferred to settle in one of the 9 least preferred states. Comparisons showed that airmen whose homes of record are in the northern tier states consider southern bases as less desirable assignment locations. However, the 50 most desirable locations for the home of record group tended to be concentrated in Colorado, Florida, and California as was found for the total sample.

Geographical - Environmental Variable Differences Between Most and Least Preferred Locations

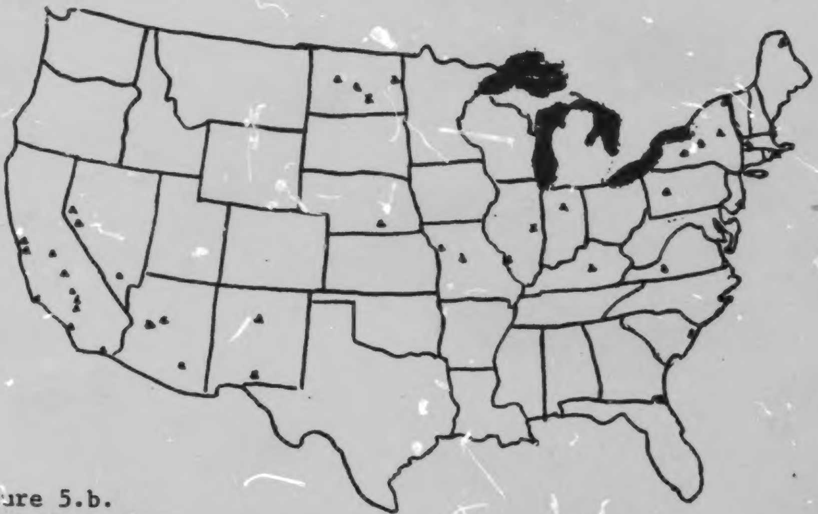
To examine certain geographical-environmental variable differences, the 30 most desirable locations and the 30 least desirable locations were identified. These most and least desirable locations were selected since they represented fairly well defined geographical areas, and appeared sufficient to reflect patterns of differences in certain geographical-environmental variables.

The data used for analysis were the numerical values of the 22 geographical-environmental variables for each of the 30 most and the 30 least desirable locations (see Figure 2). The cost of living variable which had six descriptive adjectives was recoded into a six-point scale (6 = above to 1 = much below) before proceeding with the analyses.

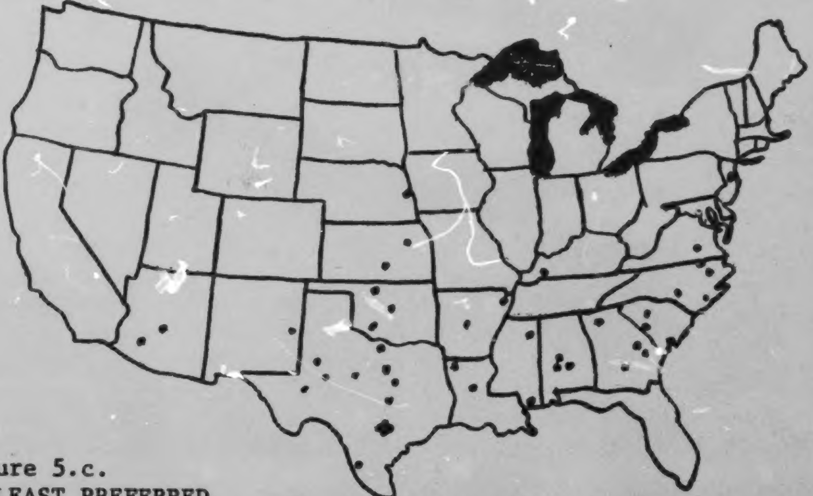
Table 2 presents the means of the 30 most and 30 least desirable locations for 19 geographical-environmental variables. Analysis of variance was used to determine the



■ Figure 5.a.
50 MOST PREFERRED

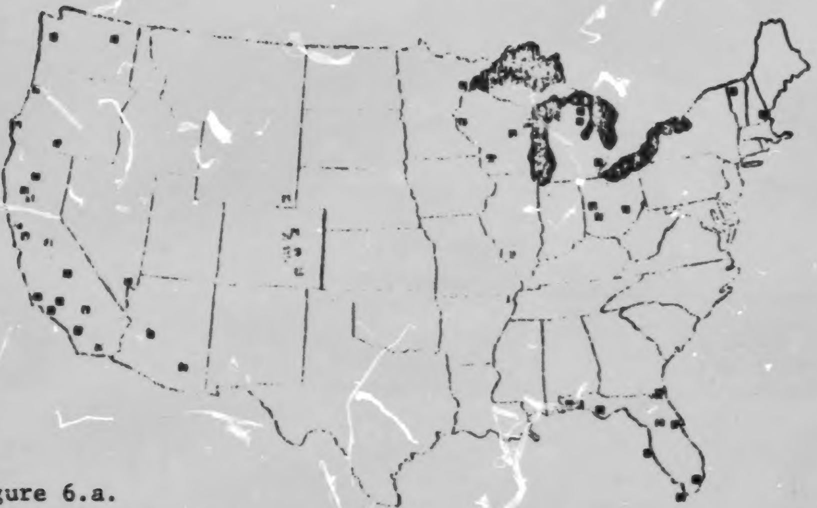


▲ Figure 5.b.
50 INTERMEDIATE

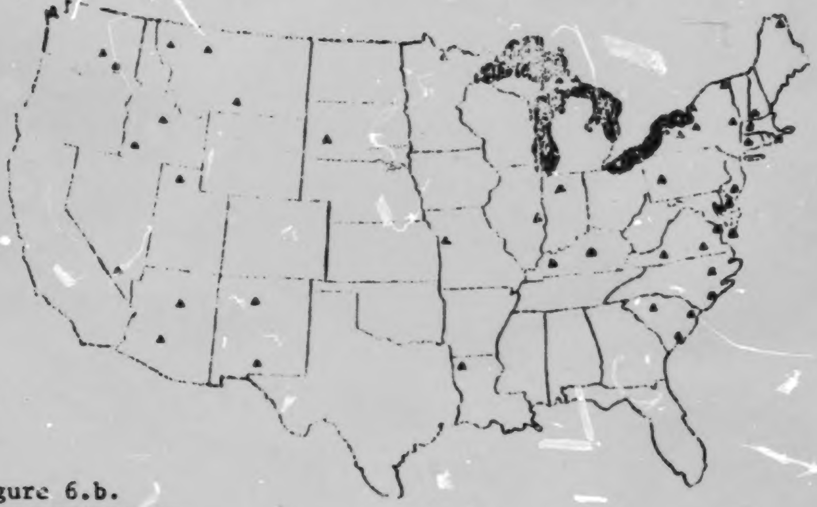


● Figure 5.c.
50 LEAST PREFERRED

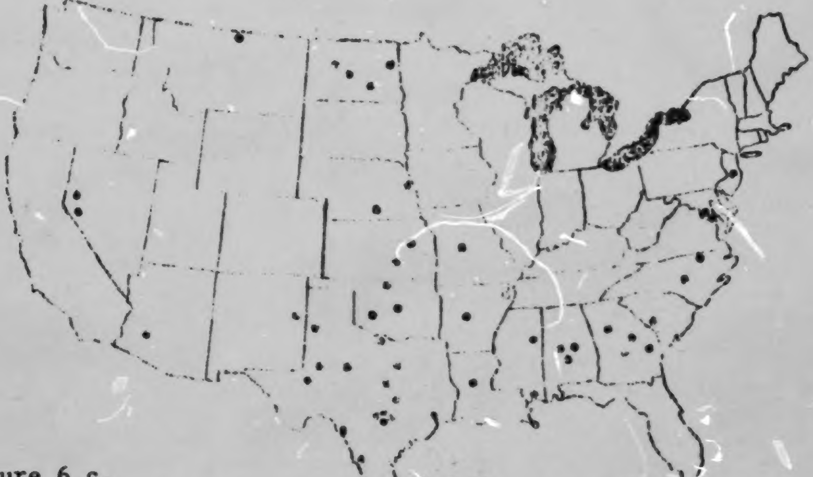
Figure 5. Plot of Fifty Most Intermediate and Least Preferred Locations of Airmen Who Prefer to Settle in Nine Least Preferred States



■ Figure 6.a.
50 MOST PREFERRED



▼ Figure 6.b.
50 INTERMEDIATE



● Figure 6.c.
50 LEAST PREFERRED

Figure 6. Plot of Fifty Most, Intermediate and Least Preferred Locations Of Airmen Whose Home of Record Is In Nine Least Preferred States

significance of the differences between the two means for each of these variables. As shown in Table 2, significant differences between most and least preferred means were found for 13 variables. A summary of these differences is as follows: (a) Most preferred locations had larger base military populations and were closer to large cities than the least desirable locations. (b) Most preferred locations were closer to oceans and deserts, yet farther from mountains and lakes/rivers than the least desirable locations. (c) The average winter (January) and Summer (July) temperatures for the less desirable locations were lower than the average temperatures for the most desirable locations. (d) The average percent of daily sunshine for both winter and summer were greater for the most preferred locations than the least preferred locations. (e) The annual snowfall was much higher for the least preferred locations, with the exception of Colorado. Of 30 most desirable locations there were five Colorado bases with relatively high snowfalls, yet the remaining 25 locations had an average of approximately zero inches. The means for the community size for the 30 most desirable locations was about four times as large as the means for the 30 least preferred.

Another question of concern was differences in availability of certain educational facilities (i.e., 2-year, 4-year, and graduate/professional colleges within one hour driving time). There were significant differences in the availability of 2-year colleges for the 30 most and least preferred locations. There were no significant differences in availability of 4-year colleges or graduate/professional schools. This situation appeared logical however. When the highest education level of the total combined sample was examined, 71 percent had 12 years of education. Probably more airmen were interested in college-entry level education (i.e., 2-year colleges), than in more advanced levels.

IV. CONCLUSIONS AND RECOMMENDATIONS

In review, an assignment location preference survey covering 150 CONUS bases or stations was completed by 4,670 airmen. Conclusions from base name ratings and other geographical-environmental variable data were as follows:

1. For the total sample, the most preferred locations tended to be concentrated in Florida, California, and Colorado. The least preferred locations were concentrated in "northern tier" states.

Table 2. Difference Between the 30 Most and Least Preferred Locations For 19 Geographical/Environmental Variables

Geo/Env Variables	Means			
	30-Most	30-Least	F*	P
Base Military Population	4,378.00	2,199.00	10.53	< .01
Community Population Near Base	212,233.33	50,350.00	2.55	NS
Miles to City ≥ 200,000	75.67	205.00	20.39	< .01
Cost of Living	3.37	3.23	.43	NS
Miles to Desert	635.30	1,061.71	5.22	< .05
Miles to Mountains	171.70	78.17	4.67	< .05
Miles to Ocean	215.70	847.33	33.66	< .01
Miles to Lakes/Rivers	27.50	8.87	8.13	< .01
Avg July Temp - High	89.23	82.70	12.30	< .01
Avg July Temp - Low	66.63	59.30	14.57	< .01
Avg % Summer Humidity	61.23	65.17	1.17	NS
Avg Days Summer Rainfall	12.50	15.50	.24	NS
Avg % Summer Sunshine	75.27	66.83	18.80	< .01
Avg Jan Temp - High	58.23	28.47	106.82	< .01
Avg Jan Temp - Low	38.90	9.97	85.85	< .01
Avg % Winter Humidity	68.60	72.48	3.35	NS
Avg Days Winter Rainfall	12.10	11.10	.39	NS
Avg % Winter Sunshine	60.03	45.33	67.68	< .01
Annual Inches Snowfall	8.43	49.30	42.49	< .01

*df = 1 and 58.

2. The strength of positive and negative valences for 150 locations was examined and revealed that the number of negative means was about five times as large as the number of positive means. There were fewer locations for which airmen have positive preferences than locations for which they have negative preferences.

3. There was a tendency for variability in preference ratings to vary as a function of level of desirability. For less desirable locations there was greater rater agreement regarding preference ratings and for more desirable locations there was less rater agreement in ratings.

4. For first-term and single airmen, many Texas bases were less desirable than the "northern tier" locations. For careerists and married airmen, many of these same Texas bases were of intermediate or high desirability and the "northern tier" locations were least preferred.

5. Home and family ties appeared to influence the relative order of preferences for the 150 locations. For airmen whose home of record was in one of nine least desirable states or who stated they would prefer to settle in one of nine such states, the least desirable locations were heavily concentrated in a southern geographical area extending from Texas through Virginia.

6. Certain geographical-environmental variable differences were found between the 30 most and the 30 least desirable locations, for the total sample. In contrast to the least desirable locations, the most desirable locations had (a) larger base military population, (b) were closer to large cities, (c) had warmer temperatures, (d) more sunshine, (e) closer to deserts and oceans, (f) lower annual snowfall (Colorado excepted), and (g) greater availability of 2-year colleges.

With regard to the present Air Force Assignment system, one suggestion from the present study is to extend the Airman Assignment Preference statement (AFR 39-11, 1973) to include undesirable as well as desirable assignment locations. While one objective of the Air Force is to assign airmen to locations for which they have positive valences, this is not always possible due to practical constraints in the assignment system. Findings suggest that when it is possible to fulfill an individual's positive preferences, procedures should be established to avoid assigning the airman to a location for which there is a strong negative valence.

REFERENCES

Air Force Regulation 39-11. *Enlisted Personnel. Airmen Assignments.* Washington, D.C.: Department of the Air Force, March 1973.

Tuttle, T.C., & Hazel, J.T. *Review and implications of job satisfaction and work motivation theories for Air Force research.* AFHRL-TR-73-56, AD-782 443. Lackland AFB, Tex.: Occupational Research Division, January 1974.

Tuttle, T.C., Brockhaus, W.L., & Hazel, J.T. *Development and feasibility test of a method to study location assignment preferences of airmen.* AFHRL-TR-74-40, AD-781 041. Lackland AFB, Tex.: Occupational Research Division, April 1975.

Tuttle, T.C., Gould, R.B., & Hazel, J.T. *Dimensions of job satisfaction: Initial development of the Air Force occupational attitude inventory.* AFHRL-TR-75-1. Lackland AFB, Tex.: Occupational Research Division, July 1974.

THE AFFECTIVE DOMAIN AND THE MILITARY SHORT COURSE

The possibility that students might undergo significant changes in attitude during military short courses (courses of 20 weeks or less) has been sparsely discussed or researched. The present study was conducted to see if significant attitude changes were possible during the $5\frac{1}{2}$ week curriculum of the U.S. Air Force Academic Instructor Course (AIC), Maxwell AFB, Alabama. The AIC mission is to teach selected Air Force and other DOD military and civilian instructors how to teach in academic settings. Due to the purpose of the AIC curriculum, it was decided to examine the two dimensions (Student-orientation and Content-orientation) obtained from the Student-Content (S-C) Teaching Inventory (Spier, 1974). A pre-test/post test design was used on four AIC classes (Total N=404). Results indicated consistent significant increase on both dimensions reflecting changes in student attitudes toward teaching. The results are discussed in relation to the large number of students that attend military short courses each year, and the attitudinal changes that can result therefrom. Implications for the military short course instructor are examined.

THE AFFECTIVE DOMAIN AND THE MILITARY SHORT COURSE

By Major Ronald G. Swanson
Academic Instructor Course
Maxwell AFB, Alabama

The short course has been an integral part of the military education system for many years, and has been used for a wide variety of purposes. Since the military services send a large number of students through the various short courses, these courses play a prominent role in military education.

While most short courses have stated objectives, few include specific affective objectives—even though attitudinal changes are implicitly expected, sought after, and desired. Some short courses do have stated affective objectives. For example, one of the course objectives for the USAF Academic Instructor Course (AIC) is "to develop attitudes appropriate for Air Force instructors." One of the course objectives for the USAF Senior NCO Academy is that each graduate "should be stimulated toward reaching his maximum potential as a senior noncommissioned officer."

Whether specifically stated or not, most military short courses appear to have at least an implied goal of changing student attitudes toward whatever the subject matter might be. This is not unusual and is in fact, in line with findings in the civilian community as well (Ringness, 1975). This does, however, pose a paradox in that while short courses at least implicitly look for attitudinal changes, students and instructors often assume that student attitudes are changed little, if any, by short course attendance. This is probably due to the fact that there is little evidence available which indicates that student attitudes can be altered significantly during short courses. In other words, though affective objectives are implied or stated, little has been done to determine if those objectives are being met.

The possibility that students might undergo significant changes in attitude during military short courses has been sparsely discussed or researched. The present study was conducted to see if significant changes in attitude were possible during the 5½ week curriculum of the U.S. Air Force Academic Instructor Course, Maxwell AFB, Alabama. The AIC mission is to teach selected Air Force and other DOD military and civilian instructors how to teach in academic settings.

Method

Subjects.

The subjects were 404 AIC students ranging in age from 19 to 54, and in rank from sergeant (E-4) to colonel (0-6). These subjects were students attending AIC during Classes 74-F and G, and 75-A, B, C and D. Ninety-seven percent were male.

Instruments.

Due to the purpose of the AIC curriculum, it was decided to examine attitudes specifically related to education. It seemed logical that a school that teaches teachers how to teach, should be concerned with student-teacher attitudes toward students and course content. An existing attitude inventory (Spier, 1974) called the S-C (Student-Content) Teaching Inventory was selected since it provided a means of determining the relative strength of two dimensions called Student Orientation (S) and Content Orientation (C).

The S-C Inventory consists of 40 items; 20 S items and 20 C items. All items are in a two-alternative, forced-choice format. S items "allow a teacher to choose between attitudes and behaviors reflecting an emphasis on sharing vs. an emphasis on personally retaining classroom authority" while C items "allow a teacher to choose between attitudes and behaviors reflecting emphasis on job activities vs. emphasis on role attributes" (Spier, 1974, p. 113).

The inventory yields a score on each dimension that can vary from 0 to 20. A sample S item follows:

It is more important for a teacher?

1. (A) To organize his course around the needs and skills of every type of student.
   (B) To maintain definite standards of classroom performance.

A sample C item follows:

26. (A) To construct fair and comprehensive examinations.
    (B) To set an example for his students.

When originally published, the S-C inventory was considered an "armchair" instrument. There had been no standardization or investigation of its statistical properties. Using the 404 inventories accumulated at AIC, the following reliability coefficients were calculated using the Kuder-Richardson 20 measure of internal item consistency for dichotomously scored items:

Student Orientation (20 items) = .75
Content Orientation (20 items) = .67

These reliabilities are quite good for an attitudinal measure and more than adequate for use in research of this kind (Nunnally, 1967).

Procedure.

During the first day or two of each AIC class the students were given the S-C Inventory. The students were not told that they would take the same inventory as a post-test during the last day or two of class. When the post-tests were administered and collected, the S and C scores for both the pre and post tests were calculated and compared via matched groups t tests and via the Wilcoxon matched pairs sign test. The primary reason for testing the significance of the differences between pre and post S-C scores by both parametric and non-parametric tests has to do with the nature of the data involved. Though good arguments could be made that the S-C scores are examples of interval data, and that parametric tests are quite robust when the assumptions underlying them are violated, it was decided to conduct a dual analysis in order to eliminate arguments about the nature of the data and the validity of the statistical tests involved.

Results

Differences between S and C scores were compared for each AIC class as well as on the total sample ($N = 404$). Only the total sample results are included here because the results for each individual class were virtually identical. Means and standard deviations for the four variables involved are presented in Table 1.

Table 1

Summary of S-C Data for Total Sample

Variable	M	SD
Pre S	14.3	3.56
Post S	15.5	3.24
Pre C	10.6	3.19
Post C	11.4	3.08

When total sample pre and post test scores were compared for Student Orientation, a significant difference resulted, $t\ (403) = 8.58$, $p < .001$. The Wilcoxon test yielded $z = 8.61$, $p < .001$. Similar results were evident when pre and post Content Orientation scores were compared, $t\ (403) = 5.79$, $p < .001$ (Wilcoxon, $z = 6.25$, $p < .001$).

To obtain some perspective on the relationship between the four variables involved, Pearson *rs* and Spearman rank order correlation coefficients were calculated. A matrix of these correlations is presented in Table 2.

Table 2

Intercorrelations Between the Four *S*-*C* Variables

Variable	1	2	3	4
Pre *S* (1)	-			
Post *S* (2)	.64* (.66*)			
Pre *C* (3)	-.06 (-.09)	-.02 (-.04)	-	
Post *C* (4)	-.05 (-.05)	.02 (.02)	.59* (.57*)	-

NOTE: Pearson *rs* are shown first followed by Spearman *rho* s in parentheses

* $p < .05$

When examining the correlations depicted in Table 2, notice that pre and post scores within each dimension are strongly related as one would expect. However, when the score relationships between dimensions are examined, little or no relationship is evident. This provides preliminary evidence that the *S* and *C* dimensions are independent. Scores on one dimension appear to have no systematic relationship to scores on the other. This indicates that the *S*-*C* Inventory is useful in that it measures two relatively independent aspects of a teacher's affect.

Discussion

The results indicate that significant changes in attitude took place during the 5½ weeks of the AIC curriculum. The differences between pre and post *S* and *C* scores were strong, and consistently reflected an increase on both dimensions for all AIC classes sampled. In other words, AIC students became significantly more concerned or oriented toward students *and* course content after the AIC short course experience.

The causes or reasons for such changes could have stemmed from many sources—the curriculum, the faculty, the physical facilities, or any number of other factors. There is abundant evidence (Himmelfarb & Eagly, 1974) that changes in attitude can occur via mere exposure, association, persuasive communication, and self-discrepant behavior such as role-playing and forced compliance—all in short periods of time. However, for the purposes of this study, reasons for the changes will not be pursued. The important point of this study is that student attitudes can be changed significantly in the short periods of time allotted to military short courses. This is not to say that such changes *will* occur, but that they can occur—and significantly so.

Hopefully these findings will cause short course administrators to pay more attention to affective objectives whether stated or implied. There is an important place in military education for feelings and attitudes. Many educators (e.g. Rogers, 1974, Krathwohl, 1964) see education and learning as much affective as cognitive or psychomotor. Insuring that students cognitively know or understand something does not mean that they are willing to use it or value it as part of their way of life. This could be of critical importance in certain military jobs where safety procedures, human reliability, and motivation are involved in attempting to make specific procedures and precautions an integral part of the individual's life style.

Though it is clear that changes in affect take place during the AIC experience, will the changes persist beyond the course? It would be very difficult to determine this at the present time. Longitudinal studies of this nature are very difficult and costly to do. Similar questions could be asked about cognitive course outcomes also—will they persist beyond the course experience and, if so, will they be used? Here again, the persistence of affective aspects of learning is evident.

If, as this study has demonstrated, student attitudes can be measured and changed to a significant degree in a short period of time, perhaps all those involved with military short courses should reconsider the nature of their jobs. Perhaps it is not just "another class coming in." Perhaps it is not that "I can't change the way I (they) feel in so and so many days or weeks." Perhaps all military short courses should have specific affective objectives and work toward determining if they're being met.

References

Himmelfarb, S. and Eagly, A. H. Readings in attitude change. New York: Wiley, 1974

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. Taxonomy of educational objectives, Handbook II: affective domain. New York: McKay, 1964.

Nunnally, J. S. Psychometric theory. New York: McGraw-Hill, 1967

Ringness, T. A. The affective domain in education. Boston: Little, Brown, 1975.

Rogers, C. R. Can learning encompass both ideas and feelings? Education, 1974, 95, 103-114.

Spier, M. S. S-C (Student-Content) Teaching Inventory. In J. W. Pfeiffer and J. E. Jones (Eds.), The 1974 annual handbook for group facilitators. LaJolla, California: University Associates, 1974.

The State of the Art in Applied Performance Testing

Thomas P. Sachse, Clearinghouse for Applied Performance Testing,
Northwest Regional Educational Laboratory
and
James R. Sanders, The Evaluation Center,
Western Michigan University

INTRODUCTION

I would like to preface my remarks this morning with a brief discussion of the measurement technique we refer to as applied performance testing, and also to the Clearinghouse for Applied Performance Testing, commonly called CAPT. The main focus of my presentation will be on the procedures we have used in obtaining applied performance tests and testing materials, as well as the results of the first year of our efforts. The intent of this discussion is to highlight currently available materials and major centers of activity in the field of applied performance testing, to provide updated information on the current status of the field, and to outline some future research and development directions for the area of applied performance testing.

Definition of Applied Performance Testing

Applied Performance Tests are instruments designed to measure performance in an actual or simulated setting. In education, performance tests focus on the measurement of performance on tasks significant to a student's life outside the school or to adult life. In vocational and military training programs, applied performance tests measure the skills or competencies a trainee must acquire for effective job performance.

Applied performance tests in education are applied in the sense that the instruments are designed to assess whether examinees can perform tasks requiring the application of classroom learning. An applied performance test measures the end point, or terminal objective, in instruction. For example, high school mathematics students who are taught computations involving percentages may be given a test which covers the material in much the same way it was taught. On the other hand, an applied performance test covering percentage could be used to assess a student's ability in computing interest on a loan-application of a skill in a practical situation. Renewed interest in applied performance testing stems from the desire of many educators (and lay persons) to direct classroom measurement toward the practical outcomes of classroom learning. Some suggest that applied performance assessment could direct curriculum development to close the gaps between classroom learning and survival skills required for quality life in a complex society.

In training settings, the distance between classroom instruction and on-the-job task requirements is greatly reduced, if not eliminated by the use of applied performance tests. Instruments developed by the Human Resources Research Organization (HumRRO) are designed to measure pre-specified objectives of a training program, yet are independent of instructional strategies employed in training. This procedure ensures measurements of proposed training outcomes, while reducing the possibility of teaching for the test. The degree of outcome specification is one main difference between training and education (training situations have specific outcomes that direct the content of instruction and measurement).

The identifying difference between applied performance and other types of tests is the degree to which the instrument approximates the reality of the situation in which the task must be performed. Applied performance tests are designed to incorporate a high degree of realism into the test situation. (Realism, fidelity and authenticity are used interchangeably in describing the degree to which these tests mimic real life situations that require this task performance.). Because authenticity to real life encounters is relative, it is sometimes difficult—perhaps impossible—to clearly distinguish between those tests that are applied performance and those that are not. There is therefore a gray area in which categorization of applied performance tests becomes very difficult.

Item stimulus and response can act as focal points for applied performance test designation and it is possible to separate most instruments on the basis of authentic reproduction of real life situations. The stimulus and response can both have either high or low fidelity. If either or both have high authenticity, the instrument may be labeled an applied performance test. Figure 1 depicts the categories of instrument that fall into this definition of applied performance testing.

An example of high stimulus authenticity/low response authenticity would be an elaborate simulation (jet simulation trainer) that requires a multiple choice response to the stimulus situation. An example of a high stimulus authenticity/high response authenticity would be a real life situation (the ability to plan and cook a nutritionally complete dinner) that calls for the correct execution of the appropriate response (cooking the meal). An example of a low stimulus authenticity/high response authenticity would be a media simulation (police training film) that calls for real life skills (trainee makes and carries out a shoot/no shoot decision, where accuracy is assessed in the former case).

Response Authenticity

Stimulus Authenticity	Low	High
Low		X
High	X	X

X = Applied Performance Task

Figure 1. Definition of Applied Performance Tests

The judging of high or low authenticity is subjective, but this matrix offers an interesting way of viewing the salient points of applied performance testing. Both descriptors, applied and performance, imply the difference between this testing technique and others. The proximity to realism in the testing situation is the distinguishing feature of applied performance testing.

CAPT Description

The Clearinghouse for Applied Performance Testing was cooperatively conceived, proposed and undertaken by a consortium of four states—Hawaii, Oregon, Pennsylvania and Washington. These member states have worked together with the Northwest Regional Educational Laboratory, to establish a mechanism for rapid access to applied performance testing materials, to disseminate information about such materials and to begin filling gaps in the availability and use of applied performance testing materials and procedures.

The initial emphasis in the collection and dissemination of applied performance testing materials was initially directed toward subject matter areas included in traditional public school setting at grade levels K-12. However, the results of early CAPT collection efforts clearly indicated that the bulk of the work in the field of applied performance testing had occurred beyond the scope of public school subject matter areas and grade levels. Extensive work had been undertaken in business, the military, vocational/occupational training and non-public school educational fields such as professional education (health fields, teacher-training, etc.) and adult education. Because of these findings collection efforts have been expanded to facilitate finding materials adaptable for public school use.

Allow me to digress for a moment to discuss the procedures involved in acquiring materials now available from CAPT. I will then return to a discussion of the results of collection activities over the past year.

PROCEDURES

July—September 1974

The CAPT Policy Board, consisting of one representative from each of the four member states, adopted a general policy directing collection efforts to be carried out by the CAPT staff. This policy dictated that collection activities would be launched through a variety of strategies. Overlap in collection endeavors would indicate full coverage of the field while new discoveries would indicate that different search methods could produce additional applied performance materials.

It is important to note that many collection activities were undertaken for reasons other than the collection of applied performance testing materials. For example, the first collection strategy was conducted to publicly announce the establishment of CAPT and to solicit information about persons and projects in the field. A publicity release was sent to national education journals and regional and state informational publications. Among those journals used to solicit information were:

Education Daily
Educational Researcher
Educational Leadership
Measurement News
Phi Delta Kappan
APA Monitor

Journal of Educational Research
Behavior Today
Educational Testing Service Recaps
Interstate Project Reporter
The Oregon Pipeline
Educational Technology

During the initial collection phase of the CAPT project, CAPT staff searched literature for tests or informational papers on applied performance testing. Because the field of applied performance testing was rather young (at least in terms of renewed interest in education), research efforts focused mainly on publications and projects from the last five years. Military sources of information, however, required more extensive research, since performance testing had been employed by the Military since World War II.

Searches of computer information bases were conducted to reveal additional sources of information. Information bases searched included:

Educational Resources Information Center (ERIC)
Research and Information Services for Education (RISE)
Vocational Education Information Network (VEIN)
National Technical Information Service (NTIS)
Psychological Abstracts

Results of the computer searches varied widely—even within the same system—depending on the search strategy and descriptions used. This variability stemmed from the fact that descriptors used within systems did not correspond to current notions of performance. This proved a complex problem; not only did descriptors fail to match our definitions (thereby making access difficult), but descriptors assigned during entry of the documents into the system were couched in a dated vocabulary.

October-December 1974

As CAPT staff conducted field surveys and literature searches, Brigham Young University (BYU) undertook a search for additional sources of information about work in the field. The BYU staff completed a literature survey as well as a field survey of applied performance testing activity. The literature survey netted 350 annotated references. Many of these references duplicated present CAPT materials; however, there were enough new references to convince CAPT staff that the external search activity had indeed been beneficial to the project.

The field survey employed a questionnaire sent to 600 individuals throughout the country. In this survey, special emphasis was placed on determining the need for and availability of instructional materials for applied performance assessment. Unfortunately, the return rate of this questionnaire was quite low--perhaps because the questionnaire was sent just prior to Christmas, 1974. A follow-up study retrieved additional data.

By the winter of 1974, many projects had been identified in the field, and the Policy Board noted approximately 30 as major centers of activity. In an effort to collect current information about new developments in the field, the CAPT Policy Board and staff visited these projects. CAPT staff developed an interview format for collecting required data. Because of variations among these centers, the interview schedule was modified to fit each situation. CAPT staff were primarily interested in obtaining data on (a) an overview of the project, (b) use of applied performance testing, (c) available measures or procedural papers on applied performance testing, and (d) knowledge about other projects using applied performance testing. These site visits proved quite beneficial. Although some projects were not as deeply involved in testing as originally believed, most had materials and references that were of great utility to CAPT and CAPT users. The following centers of activity were visited by CAPT representatives.

Visitation Sites in Washington, D.C.

- Superintendent of Documents
  Government Printing Office

- Federal Aviation Administration

- Human Resources Research Organization (HumRRO)

- Office of Research
  National Institute of Education
  U.S. Department of Health, Education and Welfare

- Test Services Center
  Personnel Research & Development Center
  U. S. Civil Service Commission

- President's Council on Physical Fitness

- U. S. Department of the Treasury
  - Internal Revenue Service (Training Division)
  - Consolidated Federal Law Enforcement Training Center

Visitation Sites in Chicago, Illinois

- Chicago Urban Skills Institute

- Center for Educational Development at the University of Illinois Medical Center

Visitation Sites at the Educational Testing Service (ETS) in Princeton, New Jersey

- Cooperative Assessment of Experiential Learning (CAEL)

- National Occupational Competency Testing Institute

- ERIC Clearinghouse on Tests, Measurement and Evaluation

- ETS Test Collection

Visitation Sites at Other Locations

- Southern Association of Schools and Colleges at Atlanta, Georgia

- U. S. Army Combat Arms Training Board at Fort Benning, Georgia

- Measurement Information Exchange (Project MIX) at Boulder, Colorado

- Cooperative Accountability Project, Colorado State Department of Education at Denver, Colorado

- National Assessment of Educational Progress (NAEP) at Denver, Colorado

- United Airlines Flight Training Center at Denver, Colorado

- Adult Performance Level Study, The University of Texas at Austin, Texas
- Research and Development Center for Vocational Education at Columbus, Ohio

January-March 1975

Ongoing activities, such as solicitations and literature searches, continued to expand CAPT holdings. Publication of the CAPT Newsletter continued on a bi-monthly basis to provide information on CAPT activities and on the status of the field. The January issue of the CAPT Newsletter contained a list of ''References Related to Applied Performance Assessment,'' subdivided into three sections according to accessibility of the materials. One section listed documents available from ERIC Microfiche collections. Another provided references to materials collected by and available from CAPT. An order form was attached for users' convenience in obtaining materials from CAPT. This list was generated before all collection activities had ceased, in order that applied performance tests and testing materials might be disseminated to educators before the school year had ended. Teachers could then begin using applied performance tests in planning their assessment activities.

On March 30, 1975, a National Conference on the Future of Applied Performance Testing was held in Washington, D. C., in conjunction with the AERA/NCME annual meetings. This conference, co-sponsored by NCME and CAPT, served to further educate interested persons in the use of this valuable assessment technique. Proceedings of this conference have been made available to those unable to participate. Conference participants were provided an opportunity to add input during small group discussion sessions, and to gain, through working sessions, a sense of direction for future research in the field. The National Conference, originally intended as a dissemination activity, was quite useful as an information collection device because applicants and participants at the conference were generous in providing information about projects and products not previously encountered by CAPT.

July-September 1975

An updated version of the document ''References Related to Applied Performance Testing'' was enclosed with the July issue of the CAPT Newsletter. The updated reference list was divided by sections indicating the availability of the materials (similar to the January version) and materials available from CAPT were categorized by subject matter for ordering ease. This document included many materials collected from January to July and provided detailed ordering information.

The September 1975 issue of the CAPT Newsletter contains a 64-page Annotated Bibliography on Applied Performance Testing. This document contains summaries of all documents available from CAPT, arranged by subject matter categories as in the Related References document. The latter two documents represent the culmination of CAPT collection efforts in the field of applied performance testing to the present.

RESULTS

Content Area Emphases

The myriad of collection activities described above portray the state of the art in applied performance testing. As one might expect, applied performance testing is well developed in subject matter areas where the product of instruction requires the same performance. Occupational fields such as carpentry, mechanics, clerical education and masonry rely on both performance tests and complementary paper and pencil tests to certify occupational competency. Professional occupations-especially the medical arts and teaching-have been very active in using performance training and testing. The Military and private industry have also used performance testing extensively.

Simulation is a well-developed facet of applied performance testing. Business and the medical arts are proficient users of simulation and gaming techniques. Simulation has some distinct advantages over actual performance testing, including reduced cost, increased sampling of behavior, reduced consequences due to error, and the possibility for variation while maintaining standardization.

Traditional public school content areas often lack applied performance testing devices. Increased interest and development in basic skills assessment, however, will soon change this; a growing concern to assess school subject matter in terms of life skills will require additional measures of an applied performance nature. But even now the Clearinghouse for Applied Performance Testing is capable of providing technical instruments for measurement of public school content areas. By assisting in the development and dissemination of new measures, CAPT is not only promoting the use of applied performance testing, but also helping to meet the growing need for validated applied performance instruments.

In searching for applied performance tests in public school content areas, substantial variance was found in the number of instruments available by content area. Applied performance tests were found for reading and mathematics at the junior high and high school levels. Although test items in reading and math were encountered for the elementary levels, most were traditional paper and pencil items rather than problem-solving or application exercises measuring mathematical competencies. Similar findings were made in psychomotor skills assessment. Applied performance tests were available for junior high and high school, but little was found for K-6 grade levels.

A number of subject matter areas were void of applied performance tests. Social sciences (including history, civics, psychology, philosophy and economics), physical sciences (including geology, geography, biology and physics) and foreign languages are content areas for which no tests of an applied performance nature were found by CAPT search activities. Certainly the latter two content areas, physical sciences and foreign languages, place heavy emphasis on performance. Therefore, applied performance test development should be expected in these content areas. The Arts (including drama, literature, and art and music forms) is another area where applied performance test development has not met its potential.

The fact that applied performance tests are not available in these public school content areas does not imply that performance measurement of student progress does not occur. Rather, performance measurement that does occur usually takes place in an informal non-standardized testing situation in which neither the examinee nor examiner are certain of the criteria (or the weightings of those criteria) upon which the performance is evaluated. One of the goals of the CAPT Project is to reduce the number of informal performance tests by highlighting positive aspects of standardized applied performance tests and by providing evaluative criteria upon which applied performance tests may be judged.

CAPT: A Resource

The CAPT Newsletter has been the vehicle by which CAPT has endeavored to extol the virtues of standardization in applied performance testing. Dr. Janet Sumida, Hawaii representative to the CAPT Project, has collected a list of evaluative criteria that can be used to evaluate applied performance tests. These criteria are available upon request from the Clearinghouse and I have brought you copies for those of you interested in reviewing or revising applied performance instruments.

The CAPT Project is continuing to be responsive to the needs of the CAPT consumer audience. During the second year of CAPT operations, evaluative criteria will be formalized into an evaluation model for judging an applied performance test. Producing means of evaluating applied performance tests is not the only effort by CAPT to advance the state of the art in applied performance testing. In addition to dissemination activities such as the CAPT Newsletter's reference lists and the Annotated Bibliography, CAPT is developing materials viewed as necessary to the expansion of the field.

One such effort is the development of instructional materials on applied performance testing. These instructional units are being produced to introduce teachers and administrators to the concepts and applications of applied performance testing for student evaluation in the classroom. An extension of the work on instructional materials will develop exportable training packages that will provide more information on the uses of applied performance testing. These packages will be pilot tested in the four member states and will be generally available after revision work has been completed.

Another CAPT product will be the Handbook on Applied Performance Testing. This handbook will include a compilation of monographs written by persons working in the field of applied performance testing. The content of this document will be more technical than the training package, yet suitable for independent study or as a resource text.

Clearinghouse representatives actively participate in conferences dealing with applied performance testing. Visitations to major centers of activity in the field are providing CAPT with new information and are establishing a mechanism for the continual sharing of new developments.

Although resources for undertaking the development of new applied performance tests are scarce, CAPT is committed to filling gaps where no instruments are currently available. To the extent possible, CAPT will begin to develop applied performance tests in the aforementioned public school subject matter areas.

CAPT staff has continued to work with agencies such as the U. S. Office of Education (USOE) and Defense Activities for Non-Traditional Educational Support (DANTES) to further the state of the art in performance testing. Tests of functional adult literacy have been collected and evaluated for USOE. These test evaluations will be used for decision making in Right to Read and other literacy programs. Tests of occupational and vocational competence have been identified and described together with descriptions of testing programs available to achieve competence in such areas, in a report to DANTES. These two adjunct projects to CAPT are an indication of the commitment the Clearinghouse has made to furthering the practice of applied performance testing.

Other Centers of Activity

Besides CAPT, there are a number of agencies working to expand current knowledge in the field of applied performance testing. The Human Resources Research Organization (HumRRO) has been designing performance tests—mainly for the military—for 22 years. Through this involvement in performance testing, HumRRO has produced valuable procedural papers on developing performance tests and on conducting task analyses. The Human Resources Research Organization has seven divisions located throughout the country. Personnel including: Saul Lavisky at HumRRO's main office in Alexandria, Virginia; Thomas Sticht at the Monterey, California Division; William Osborne at the Fort Knox, Kentucky Division, have been involved with and supportive of the CAPT Project.

The Center for Educational Development (CED) performs various support functions for the University of Illinois Medical Center, including designing assessment procedures for the allied health professions. In this area of support, CED is currently concerned with:

1. Working with School of Medicine faculty to design tests that will assess the accuracy of clinical judgments, including written simulations, mannequin simulations and checklists for the performance of specific tasks.

2. Working with specialty boards to design tests to certify competence in medical specialties.

3. Research and development efforts in the area of self-assessment techniques for physicians in practice.

Drs. Christine McGuire and Hulda Grobman have given valuable assistance to the CAPT Project, especially in planning the National Conference on the Future of Applied Performance Testing. They have also added to the growth of the field in simulation measurement through instruments produced and papers published at the Center.

The Cooperative Assessment of Experiential Learning (CAEL) located at Educational Testing Service in Princeton, New Jersey conducts activities that fall into three major phases. The first phase involves research and development efforts to produce techniques or processes for assessing experiential learning. The second phase concerns the validation of products developed in the first phase. Various member institutions have volunteered to field test products of the development phase. The third phase is the dissemination and application of validated products. This phase will include the training of individuals in the use of the developed and validated procedures. Tangible output to date includes three major publications:

1. ''CAEL Resource Book'' includes references to persons/projects working in the field of experiential learning, and a bibliography of references on experiential learning.

2. ''Current Practices in the Assessment of Experiential Learning,'' a working paper on assessment techniques currently in use in various institutions of higher education.

3. ''A Compendium of Assessment Techniques,'' another working paper. Dr. John R. Valley directs this project and has been the main CAPT contact. Information sharing with this organization has been most beneficial to the CAPT consumer audience.

The National Occupational Competency Institute (NOCTI) also located at Educational Testing Service in Princeton, New Jersey was originally known as the National Occupational Competency Project. The project had three main objectives:

1. To establish a consortium of states needing occupational competency tests designed to certify teachers of that particular occupation.

2. To determine the state-of-the-art in occupational competency testing—it was found that only a paucity of tests existed and

in many cases state departments of education were duplicating other efforts in test development.

3. To develop needed tests and associated testing materials.

Since becoming an institute, NOCTI has explored the feasibility of administering national occupational competency tests; 1200 examinations are now administered yearly to future teachers in each of 24 occupational areas. These exams are used for three main purposes:

1. To obtain data for selection in teacher education programs.

2. To provide proficiency credits by examination.

3. To certify competency in an occupation.

Dr. Adolf Panitz, Director of the original project as well as the present institute, has assisted CAPT through personal dialogue and directing CAPT representatives to other major centers of activity and documents in the field.

The Combat Arms Training Board (CATB) in Fort Benning, Georgia is responsible for developing instructional systems and complementary materials (including performance tests) for all military occupation specialties. Within this framework, CATB is developing an Instructional Systems Development (ISD) Model and several manuals (presently under development) that will be used in training. In addition, CATB is developing many performance-based tests in connection with performance-based instruction programs. Although some of these materials are available at CATB, most are located at training sites. The goal of CATB is to provide in various—eventually all—Army occupations.

The objectives of the Adult Performance Level (APL) Project located at the University of Texas at Austin, under the direction of Dr. Norvelle Northcutt, are to (1) develop a meaningful, accountable description of adult functional literacy for the United States, and (2) create techniques for assessing functional literacy which are useful in national and state policy planning.

In three years of research, the APL Project has developed a taxonomy which conceptualizes adult functional literacy as the application of a set of skills (communication, computation, problem solving, and interpersonal relations) to a set of ''general knowledge areas'' which are particularly important to success in American society as operationally defined by income, education and occupation. The general knowledge areas are consumer economics, occupationally related knowledge, health, community resources and government and law.

Adult functional literacy is described in terms of ''objectives'' which relate to these skills and general knowledge areas. Adult functional literacy is seen by APL as mastery of these objectives to maximize the likelihood of ''success'' in terms of being better educated, making more money, and holding better jobs. The more objectives an adult masters, the higher his/her level of literacy.

The APL Project has developed an assessment technique involving random samples of adults throughout the country who belong to one of several levels of literacy. These data, in combination with the APL operational definition of ''success,'' provide a well-defined estimate of the total number of adults in the country who belong to each level of literacy.

The Ohio State University Center for Vocational Education houses the Task Inventory Exchange (TIE) under the direction of Robert Taylor. The Task Inventory Exchange conducts task inventories or analyses; procedures that delineate all activities performed by workers on a particular job. Task analyses are of major utility to curriculum developers and performance test developers. These task analyses are used in identifying content for instruction or for measurement devices for applied performance tests.

In addition to individual projects and agencies such as those listed above, many public school education units from classrooms to state departments of education are beginning work in the field of applied performance testing. Movements toward competency-based measurement in which applied performance testing is most useful are observed at state levels, particularly in New York, Oregon, and California. Many districts in these and other states are beginning to stress real life outcomes of classroom learning. These local efforts will turn to agencies mentioned above for technical assistance in planning and implementing applied performance assessment. Education journals are not seen presently as a source of current information on the theoretical and practical implications of this measurement technique.

Other Sources

Although no single journal can be identified as a source of information on applied performance testing, occasionally articles appear on this measurement technique. The following are journals that print papers related to applied performance testing:

Agricultural Education Magazine
Child Development
Educational and Psychological Measurement
Educational Leadership
Educational Technology
Journal of Educational Measurement
Journal of Experiential Education
Journal of Industrial Teacher Education
Journal of Teacher Education
Mathematics Teacher

Perceptual and Motor Skills
Psychological Reports
Reading Research Quarterly
School Review
Teacher Educator
Training

Many additional sources of information on applied performance testing have been mentioned earlier in the procedures section. Certainly, ERIC and similar computer-based information retrieval systems are valuable information bases for general inquiries. (Access is limited, however, by the inadequacy of the descriptors used.) CAPT products, available upon request, are most directly relevant to the field of applied performance testing, as are documents available from major centers of activity mentioned above. Occasionally, chapters in measurement texts deal with performance measurement or product evaluation. These overviews are usually adequate as introductory materials.

DISCUSSION

Interest in, and use of, applied performance testing as a measurement technique has expanded over the past few years. Increased emphasis in public school education on ''real life'' application of classroom learning has been the main factor in the increased awareness of applied performance testing. This emphasis becomes clear as districts and states move toward competency-based instruction and measurement.

A number of states and districts are currently moving to competency-based measurement for promotion or graduation requirements. Oregon law now mandates the assessment of minimum competencies in a variety of subject matter areas as a prerequisite to high school graduation. The State of Washington has initiated a program allowing a student ''early exit'' from high school after certification that he or she has satisfactorily demonstrated competence in a number of required subject matter areas. Hawaii has implemented a statewide assessment program that employs competency-based measurement devices.

There are many reasons why school people are adopting competency-based educational programs as a substitute for (but oftentimes in addition to) the traditional methods of promotion—that is, completion of coursework, or school attendance. For example, the increased emphasis on the specification of educational objectives in instructional programs has encouraged adoption of competency-based programs. Behavioral or performance objectives include a delineation of the specific knowledge or skill to be demonstrated, the manner in which the task will be undertaken by the student and the precise level of performance which will be considered satisfactory in completing the objective. Detailed specification of educational objectives allows increased accuracy in the measurement of educational goals.

The interest in using competency-based instruments greatly preceded theoretical refinements and development of these measures in education. Development began at the classroom level, and gradually proceeded to the state level in education. The military and private industry (to a lesser degree) have prepared themselves via a slower evolutionary process.

The rationale for the military and private industry's interest, acceptance and use of applied performance testing relates to another major reason for the educational community's growing interest and use of this technique. As had been stated earlier, applied performance testing measures performance of some task significant to life outside the instructional setting. This technique requires performance in a simulated or actual setting. Therefore, those agencies whose instruction enables a student to perform an important life task view applied performance testing as a measurement technique consistent with outcomes of instruction. The military and business communities train personnel to perform a specific job; and measurement is tied to the actual job performance. Education, which has in the past been largely preparatory to generally defined life goals outside the school, is now committing itself to the acquisition of life skills. Because of the new commitment to teach skills which are viewed as valuable for adults in a complex society, education has found applied performance testing a valid measurement technique.

Adoption of this technique within education has, however, varied substantially. Applied performance testing is seen as valuable in those content areas in which the desired product of instruction is performance. Examples of content areas in which applied performance testing is relevant include: vocational/occupational education, physical education, the arts, and foreign languages. It is evident that even within the content areas for which applied performance testing is relevant, the relative emphasis placed on development of standardized measures has differed widely.

Two additional factors relate to the development of formalized applied performance tests. One factor is the degree to which a move toward standardization is seen as important or feasible. In vocational and physical education, the emphasis on standardization has resulted in many formalized applied performance tests. In the arts and foreign language education, less emphasis has been placed on formalizing applied performance tests. This is most likely due to the feasibility of standardizing aesthetic performances in the arts, and to interest in or need for such tests in the foreign languages. At the present time it is difficult to justify the formalization of applied performance tests in all content areas.

Another factor yet to be explored regarding development of applied performance tests is grade level. Because of the preparatory nature of elementary education, less subject matter within that level is relevant to applied performance testing. Surely reading and math skills are not yet developed to the point where application of life skills is necessary for student evaluation. The scope of most subject matter areas is reduced to lower levels of Bloom's and Krathwohl's taxonomies. Therefore, applied performance testing is less applicable in lower grade levels where performances are not part of the learning outcomes. However, this is not to say applied performance testing is never applicable at the elementary level. Tying one's shoe or buttoning an overcoat are examples of tasks best measured in an applied performance setting.

The rationale behind development of applied performance testing is multi-faceted indeed. We have examined three main factors that contribute to the use of applied performance tests: (1) relevance of the technique to measure expected outcomes in the content area, (2) desire for or feasibility of standardization of informal performance measures currently in use, and (3) grade level or taxonomic complexity of acquired skills.

Another very important consideration is the treatment of applied performance testing as but one small part of the entire testing spectrum available to those involved in instruction and evaluation. It is not a panacea for testing problems in training and education, it is not a replacement for the many highly developed technical tools we are now using; and it is certainly not a synonym for a new curriculum movement in our schools. It is a way to collect information about task performance that is useful in certain training and education settings and should be treated as simply that.

Factors such as cost, administration and scoring time, the availability of applied performance tests and compatibility of applied performance testing with the present system for student evaluation must be considered in planning for the use of applied performance tests. Although these factors tend to limit the practicability of applied performance testing, under certain conditions, they do not negate its utility. Current research and development should help to offset negative aspects of such factors. Group administered applied performance tests and synthetic performance tests and simulations can open the way for feasible applied performance testing without compromising the utility of this technique. As public support for this type of measurement grows, CAPT and many other agencies will expand the state-of-the-art in applied performance testing and thereby close the gap between the demand for and availability of applied performance tests.

SAMPLING AND SIMULATION IN TESTS OF JOB PERFORMANCE

by

Elaine N. Taylor, Robert Vineberg and William C. Osborn
Human Resources Research Organization

Two fundamental problems in constructing tests to assess job performance are: 1) how should tasks be sampled to provide an adequate representation of a job and 2) how should performance be tested when the conditions of the real task cannot be entirely reproduced? Decisions about sampling and simulation inevitably are shaped as a joint consequence of both theoretical and practical considerations.

What are the dimensions that are likely to provide the most valid sample for assessing job proficiency? If a classification of the behavioral dimensions underlying the performance of tasks were available, this question could readily be answered. Unfortunately, no really adequate understanding or agreed upon classification of the underlying demands of jobs exists.

Thus, the person planning the sampling strategy must turn to fairly gross characteristics of tasks with the expectation that they are related to the underlying demands. Dimensions commonly used are task difficulty, functional descriptions of job performance, frequency of task performance, and estimates of task criticality.

In selecting dimensions for sampling, functional area and task difficulty are to be preferred in that they are probably more closely related to the underlying demands of jobs. The relative occurrence or frequency of tasks and the criticality of tasks, while having some impact upon performance, reflect system requirements more closely than underlying behavioral requirements.

Practical considerations almost always influence the sampling strategy to be used. How much time can be used? How much money can be spent? To illustrate the influence that such practical constraints have upon sampling, let us consider the Army Skill Qualification Testing program as an example. While decisions about the Skill Qualification Testing program are subject to change, present plans limit testing time for an individual examinee to eight hours. Four of these hours are to be used in what the Army calls a hands-on component and the other four are to be used in a written component.

The ultimate implication of this practical consideration is that a skilled--non-skilled dimension must be included in the sampling plan. How do we arrive at this?

In order to understand the necessity for differentiating between skilled and non-skilled tasks we have to carefully examine what is meant by written and hand-on tests. Let us approach these questions by introducing two other terms that you are undoubtedly also familiar with--performance-oriented testing and knowledge-oriented testing. It is with these latter terms that we can best clarify the meaning and application of written and hands-on tests.

Performance-oriented tests are those in which a person is required to respond in a way that is highly similar to the way he would respond in actually performing the task on the job. His behaviors, the things he does, are very like the things he does in the actual task. Likewise the cues in the test are very similar to the cues found in the real task. The things the person sees, feels, hears, or otherwise responds to, remain essentially the same.

When some variety of simulation is used in a performance-oriented test the cues and/or the responses may be somewhat degraded, of course, from those present in the actual task but they will remain largely the same. For example, since a Military Policeman normally controls traffic by responding to moving vehicles that he sees, a performance-oriented test of traffic control requires some representation of vehicles that are actually moving. In a simulation an examinee, rather than seeing actual traffic directly, might respond to traffic on a film or video tape.

By the same token the responses in a performance-oriented test remain essentially the same. For example, performance-oriented tests of combat marksmanship against live targets are now being tried by substituting a laser weapon for an actual rifle. Here the soldier still tracks the target and squeezes a trigger even though his total response has changed somewhat in that he no longer has to be set to absorb a recoil.

The key defining property of a performance-oriented test is that the cues and responses, even in a simulation, remain in essentially the same form as in the actual task.

Knowledge-oriented tests are those in which a soldier is required to recall or recognize the information that is necessary to perform a task. The response a soldier makes in a knowledge-oriented test is always different in form from the response he makes in the actual task.

For example, a soldier might be asked to select the correct answer to a written question about energizing a radio or where to place a tourniquet after being given information about the location of a wound. In either case, the soldier does not demonstrate the actual response required in the task or really even an approximation of the response.

Another type of knowledge-oriented test might require a soldier to indicate whether the performance of some task he views on video tape is correct or incorrect.

Or he might be asked to arrange, in proper sequence, a scrambled set of pictures of the steps of a procedural task being performed. Or he might be asked to arrange written statements about the steps to be performed. In all of these instances he is being asked to display knowledge about performance. He is not being asked to demonstrate performance either in a reproduction of an actual task or in a simulation.

In knowledge-oriented tests the cues that the soldier responds to may be quite similar to those encountered in the real task or they may be very different. The key defining property of a knowledge-oriented test is that the responses to be made are always different.

Now, let us return to the terms we began with: hands-on components and written components of Skill Qualification Test

We can clearly see that hands-on components are made up of, in fact are synonymous with, performance-oriented tests.

Written components, however, can consist of either knowledge-oriented tests or performance-oriented tests.

If a task actually involves the use of written materials and/or the production of written materials, testing of performance in this task would appropriately use the written materials or facsimilies of them. The examinees responses would take the same form as in actual task performance, and the test though written, would be a performance-oriented test.

For example, an examinee might be asked to review a set of personnel records of a number of men and to identify those who are eligible for overseas levy. In making these determinations, the examinee would be provided with a set of appropriate regulations or excerpts from the relevant regulations and would proceed in making decisions in the same manner as he would on the job. The test, though dealing with written materials, would be performance-oriented.

If, on the other hand, the test consisted of multiple choice questions about the information that makes a man ineligible for overseas levy, the written test would be knowledge-oriented.

Clearly, the hands-on component, as used by the Army in its Skill Qualification Testing program, will be performance-oriented; the written component may be either performance-oriented or knowledge-oriented.

Now, when should performance-oriented tests be used and when should knowledge-oriented tests be used?

Performance-oriented tests are best used for assessing performance of tasks which require skilled behaviors.

By skilled behaviors we mean those that require practice or rehearsal during learning to provide for the discovery, or adjustment or refinement of the behaviors themselves. Often practice is necessary if the standard of task performance includes criteria such as speed, precision, or stability of performance under stress.

Examples of skilled behaviors are: aiming a rifle at a moving target, manipulating materials by a crane, or hitting a golf ball to the place you want it to go. (In addition, without going into any detail, we would like to add that cognitive tasks are skilled even though they will often be tested in the written component of the Skill Qualification Test as performance-oriented tests.)

Performance-oriented tests are necessary to assess skilled behavior since a display of knowledge, only, about task performance is not sufficient to infer that a person can perform a skilled task.

Knowledge-oriented tests, on the other hand, are appropriately used where a task is non-skilled. If it can be assumed that knowledge of procedures is sufficient to assure that a task can be performed, the task is non-skilled. Examples of tasks that are non-skilled are changing a tire, entering a radio net, or keeping your score in golf.

In deciding whether a performance-oriented test is essential, it is the skill requirement of the task that must be considered. If skilled behaviors are not involved, knowledge-oriented tests can be used.

Now, where has this discussion led us with regard to a rule about sampling? . . . . Practical considerations will almost always require that there be some knowledge-oriented testing. Therefore, decisions about sampling must be preceded by an analysis of the job into skilled and non-skilled tasks in order to identify tasks for which knowledge-oriented testing is appropriate. In short, the practical requirement that there be some knowledge-oriented testing dictates that the skilled--non-skilled dimension be included in the sampling matrix.

Finally, where has this discussion led us with regard to a rule about simulation? . . . . Any simulation of tasks involving skilled behaviors must maintain the same form of cues and responses as found in the real task.

A Criterion - Referenced Pilot Performance Measurement System

Fuller, John H., Waag, Wayne L., Eddowes, Edward E.
and Fuller, Robert R.

Flying Training Division
Air Force Human Resources Laboratory
Williams AFB, AZ 85224

Introduction

The Flying Training Division of the Air Force Human Resources Laboratory conducts research on design and utilization of future flight simulators, with emphasis on undergraduate pilot training. The majority of this research utilizes the Advanced Simulator for Undergraduates Pilot Training (ASUPT). This device simulates the T-37 jet trainer. It incorporates a six degree of freedom motion system, a pneumatic g-seat, wide field-of-view computer image generated visual system and various advanced instructional features.

Most research on simulator configuration, training techniques or advanced feature utilization ultimately require evaluation of student performance as related to these various conditions. Although instructor pilots normally evaluate student performance, there are advantages to an automated measurement system that measures performance using parameters available in the simulator computer subsystem. These advantages include:

1. Automatic measurement can be more objective and standardized.
2. Automatic measures can detect subtle differences in performance not always observable by the instructor.
3. Certain maneuvers are difficult to evaluate in the simulator due to the characteristics of the visual system.
4. Automatic measures are more convenient when numerous repetitions are required, as in certain multi-factorial research designs.

Measurement System Requirements

One of the salient characteristics of flying is that it is criterion-directed. For the execution of any maneuver or sequence of maneuvers there are definable objectives which must be accomplished. The degree to which these objectives are met represents an adequate description of performance. In other words, a criterion-referenced approach to pilot performance forms the basis for the present effort. Following such an approach, it is apparent that the definition of criterion objectives is of foremost importance. Within the context of measurement development for ASUPT, the critical question is whether these criterion objectives can be stated in terms of parameters available within the simulator? In other words, can behavioral objectives be defined in terms of the state of the simulated aircraft and control inputs of the pilot?

Aside from the requirement to define performance in terms of observable behavioral objectives, there are other constraints which are applicable. The first is parsimony in the selection of simulator parameters to be sampled. Criterion objectives are to be defined using as few parameters as possible. Furthermore, they are to be sampled and analyzed on a real-time basis so that the resulting measurement and feedback are immediate. Since measurement will be an integral part of training students in ASUPT, it is also necessary that the resulting output be meaningful and easily interpreted by other instructors and students.

In summary, a criterion referenced approach to measurement system development is to be pursued within the constraints of the following requirements:

1. Measures will assess the degree to which the criterion-objectives are met.
2. Measures will reflect only the most salient characteristics of performance.
3. Measures will be meaningful and interpretable to the user -- the student and instructor pilot.
4. Measures will be generated on a real-time basis so that feedback is immediate.

Measurement Development for ASUPT

Although the criterion-referenced approach represents the rationale for present developmental efforts, a further set of assumptions have guided implementation on ASUPT.

1. Measurement system development parallels skill acquisition in the student pilot. Since student pilots acquire flying skills in a hierarchical fashion, measurement development should proceed in a similar manner. Simply stated, a building block approach is assumed in which measurements for basic flying skills are the first to be implemented.

2. The focus of the measurement system is the individual flight maneuvers. This level of measurement seems to be most consistent with present flying training syllabuses.

3. Maneuvers can be conceptualized as integrated sequences of steady states and transitions. The fundamental flight attitudes plus transitions from one attitude to another form the conceptual segments for most maneuvers. It is necessary to define behavioral objectives for each segment since they are likely to differ from one to another.

4. Maneuver performance can be evaluated by two sets of parameters -- those reflecting the state of the aircraft and those reflecting the control inputs of the pilot. Superior performance is assumed to manifest three characteristics: (a) accomplishing the criterion objectives as defined by the aircraft state parameters; (b) avoiding excessive rates and acceleration forces so that the maneuver is executed smoothly; and (c) accomplishing these objectives with the least amount of effort; that is, by minimizing control inputs.

5. The evaluation of performance for a given parameter involves a comparison of the obtained value with some ideal value. For parameters reflecting aircraft state, the deviation from the ideal provides an index of error. Since the ideal is seldom attained, it is more realistic to define acceptable performance in terms of an empirically determined tolerance band about the ideal value. For parameters reflecting smoothness and control inputs, the adapted ideal represents a performance level characteristic of the highly experienced pilot.

6. The implementation of the measurement system requires four phases of development: (a) definition of criterion objective in terms of a candidate set of simulator parameters; (b) evaluation of the proposed set of measures for the purpose of validation and simplification; (c) specification of criterion performance by requiring experience instructor pilots to fly the maneuver in question; and (d) the collection of normative data using students as they progress through the training program.

Preliminary Development and Evaluation

In keeping with the approach outlined above, the first maneuver for which measures were developed was straight-and-level flight. Using the pre-programming capability in ASUPT, a simple test scenario was developed: (1) initialize the simulator to 15,000 feet, 160 knots, heading 090 degrees; (2) unfreeze the simulator and allow 10 seconds for the pilot to "settle down", (3) sample selected parameters for one minute; and (4) freeze the simulator. The parameters sampled were altitude, airspeed, heading, stick movement, throttle movement, elevator stick force, pitch rate, pitch acceleration, roll rate, roll accelerations, vertical velocity, and vertical acceleration. Mean deviation and root-mean-square (RMS) deviations were computed for altitude, airspeed, heading, and stick force. RMS scores were computed for the remaining parameters.

A seven-point rating form was developed for use by instructor pilots who were to provide qualitative evaluations. Five items were to be rated: (1) altitude control; (2) airspeed control; (3) heading control; (4) overall level of performance; and (5) smoothness. For each evaluation, two raters were used -- one inside the cockpit and another at the console who only observed the repeater instruments. In this manner, an estimate of inter-rater reliability could be computed.

Subjects for the initial evaluation were 12 AFHRL/FT employees and one student pilot. Piloting experience ranged from zero to 8000+ hours. Each subject was given a two-minute practice/instruction period followed by five one-minute runs. Three instructor pilots were used as raters and were randomly alternated between the cockpit and console. All missions were flown using full motion and G-seat. Of the total 65 runs, one was lost due to system failure.

First, correlations between raters were computed. All coefficients were found to be high (.85 - .93) except for the smoothness rating (r=.65). This is most likely due to availability of motion cues for the rater inside the cockpit. In any case, the high intercorrelations suggest the ratings to be a reliable criterion against which to validate the objective measures.

The next step was to determine whether the objectively derived measures would reliably predict the instructor ratings. The adopted criteria were the overall rating and smoothness rating. Correlations were computed between the objectively derived measures of performance and these instructor evaluations for both the cockpit and console raters. The results suggest several things. First, the correlations between the measures and instructor evaluations are fairly consistent for both the cockpit and console ratings. Second, and most importantly, substantial relationships exist between a number of the objectively derived measures and the IP's subjective evaluations. Third, the measure RMS vertical velocity predicted both criteria quite well. These findings were highly encouraging -- suggesting that these objectively derived measures do relate to the instructor's evaluation of performance. Using a forward selection multiple regression procedure, linear combination of variable subsets were selected which were predictive of the criterion. A multiple

R of .90 was obtained for the prediction of the overall rating.

The results of the first study were highly encouraging and seemed to warrant the following conclusions: First, instructor pilots are highly consistent in their evaluations of performance. Consequently, it is possible to use such evaluations as one criterion against which to validate objective measures of performance. Second, the objective measures of performance developed for straight and level will reliably predict instructor evaluations. The demonstration of such predictive validity suggested the approach taken to be a reasonable one. To further evaluate the proposed system, a second study was undertaken.

Scenarios were developed and the software written for the following maneuvers: change of airspeed; constant airspeed climbs/descents; rate climbs/descents; and the steep turn. Similar rating forms were developed for each maneuver and the simultaneous cockpit/console evaluation procedures followed. Four T-37 instructors were the raters -- two alternating at the console and the other two alternating in the cockpit. Ten subjects were used in the second study, again representing a wide range of skills. Four student pilots in T-37 training, three T-37 IPs and three civilians were included. Each subject flew the following set of maneuvers: three airspeed changes; one constant airspeed climb; one constant airspeed descent; one rate climb; one rate descent; and three steep turns. For each climb/descent a level-off to altitude was required.

Inter-rater correlations were computed for each maneuver. The data for climbs and descents were pooled. Overall inter-rater correlations were computed for categories which were rated for all maneuvers. The data indicates substantial agreement among the raters, especially for the overall and smoothness ratings, even though the values were somewhat less than obtained for straight-and-level. Several possible reasons for the lowered inter-rater correlations should be mentioned. First, the maneuvers in the second study were of increased complexity. Since these maneuvers require several transitions in addition to a steady state condition, the instructor's job of monitoring all the relevant parameters is increased. Likewise, the performance of transitions from one steady state to another increases the number of cues available to the rater within the cockpit. A second possibility concerns rater bias. It is possible that different subjective criteria were used in ratings of the T-37 IPs as opposed to students. An examination of the ratings of one of the IP's performance records yielded large discrepancies between the cockpit and console ratings. The objective measures appeared to agree with the cockpit ratings in that the performance was quite good. However, according to the console rater, the performance was considered to be unsatisfactory. Such data strongly suggest the possibility of rater bias.

Correlations between each of the objective measures and the overall and smoothness ratings for each maneuver were computed. In this case, the ratings from the IP in the cockpit were used as the criteria. A perusal of the data warrants several conclusions. As expected, parameters reflecting performance of the criterion objectives were most related to the ratings. Likewise, the RMS values about the ideal were correlated more highly than mean deviations. The only exception was RMS bank error. In this case, an error in the computing software was discovered, thereby invalidating the resulting measures.

A forward selection regression analysis was computed in an attempt to develop prediction equations for the overall steep turn rating. The steep turn was selected since it represented the most difficult of the maneuvers tested. The initial subset of seven variables selected by the procedure yielded a multiple R of .8820. This equation was then used to predict the console ratings. The obtained correlation between predicted and observed console ratings was .9137. Although not verified, it seems likely that similar sets of prediction equations could have been developed for the other maneuvers. In any case, it is certainly clear that the objective measures developed are highly predictive of instructor ratings, thereby demonstrating the validity of the present approach to measurement.

A further set of analyses were computed in order to relate the objective measures to the instructor evaluations. Performance records for all maneuvers were pooled and placed into groups according to the evaluation of the cockpit instructor. Four groups were defined according to the grade assigned to U, F, G, or E. Descriptive statistics were then computed for each of the four groups. Measures for rate of climb and bank were deleted due to the small number of cases within each of the groups. An examination of the data warrants several conclusions. First, it is apparent that there are clearly defined trends for a number of the objective measures across the four groups. There are clear cut decreases in root mean square for air-speed, heading, and altitude as a function of the subjective ratings. Likewise, there are decreases in the variability across subjects. Consequently, lowered ratings are characterized by increasing within-subject error as well as increasing between-subject variability. Such data reflect the fact that there is one way to execute the maneuver correctly, but many ways to commit errors and therefore receive a lower evaluation.

The results also verified the assumption that superior maneuver performance involves the minimization of control inputs. Performances rated excellent were those which minimized the amount of stick movements and also minimized stick deviations from the null position. Furthermore, superior performance was also characterized by minimum amount of control force, or, in other words, the efficient use of trim. While stick inputs were found to be related to rated performance, throttle movements were not. Of the proposed measures of smoothness only pitch rate, vertical velocity and vertical acceleration were related to rated performance.

Again, the pattern of decreasing means and variances was found.

A further requirement of the measurement system is that it reliably discriminates between pilots of different experience levels. To test this, subjects were placed into one of two categories - high vs low experience - and descriptive statistics computed. The results indicated that a substantial number of the objective measures will differentiate between the two groups. Generally, the inexperienced group was characterized by higher error scores and much greater variability. For the steady state parameters, the RMS error scores were more discriminative than the mean error scores.

The results of these two evaluation studies warrant a number of conclusions. First, the data suggest the approach taken to be a viable one. The assumptions concerning superior maneuver performance have been coorborated by the data. Experienced pilots, and likewise those performances rated excellent, were characterized by: (1) meeting the criterion objectives in terms of minimizing aircraft state errors; (2) minimizing rates and accelerations -- at least in the pitch and Z axes; and (3) minimizing control inputs. In other words, superior performance involves getting the job done, doing it smoothly, and with a least amount of effort.

Second, the data indicate that instructor evaluations are a usable criterion for future measurement system development. The relatively high levels of agreement between the cockpit and console ratings are particularly encouraging since the availability of cues was radically different for each rater. The console rater only had access to the repeater instruments while the instructor in the cockpit could observe the students' behavior in addition to the flight instruments. Furthermore, kinesthetic cues were also available to the cockpit rater. The importance of these different cue sources in instructor evaluations is an area which should be addressed in future research studies.

Third, the objectively derived measures were shown to possess a certain degree of validity. Significant correlations between these measures and instructor evaluations were obtained. Furthermore, the measures were shown to discriminate between pilots of different experience levels. These results are most encouraging and convincingly demonstrate the fruitfulness of the present approach.

Fourth, the results provide some insight into the manner in which instructors assign grades. A hierarchical model seems most consistent with the data. To receive an excellent (E) rating, errors on the critical parameters must all be low. As the quality of the rating decreases, the potential for different errors increase. In fact, it is possible to commit error involving one parameter, control the others quite well, and still receive a low evaluation. This suggests that both the number and degree of error are important. The investigation of instructors grading strategies is another prime area for future research.

Fifth, the results provide preliminary data concerning the simplification of the present set of parameters. As expected parameters reflecting performance of the criterion objectives yielded the highest validity coefficients. However, measures of mean error were not as effective as root-mean-square error. Furthermore, a number of the proposed measures of smoothness did not produce any significant relationships. Roll rate, roll acceleration, and pitch acceleration were not related to either instructor ratings or experience level. Likewise, throttle movements were not found to be important.

Recent Developmental Efforts

The preliminary development and evaluation yielded evidence suggesting both the feasibility and validity of the criterion-referenced approach to the measurement of pilot performance. As a result, a concentrated effort was undertaken for the implementation of automated measurement for all phases of T-37 training. These included representative maneuvers from the basic instruments, basic contact, advanced instruments, aerobatics, and formation phases of flying. To date, most maneuvers measuring basic instrument and contact, takeoff/landing, and advanced instrument skills have been completed. A list of maneuvers for which automated objective measures are available is presented in Table 1.

During the early implementation of these measurement scenarios, it became apparent that two types of outputs were required: (1) measures of interest to the researcher; and (2) measures which would provide immediate feedback in a meaningful and interpretable format to the student and instructor. Measures based on state parameter deviations, rates and accelerations reflecting smoothness, and control input data reflecting pilot behavior seemed to be appropriate for the researcher. On the other hand, measures derived from tolerance bands on selected state parameters as well as discrete information reflecting aircraft configuration and unsafe operating conditions seemed most appropriate for student and instructor feedback. Initially, arbitrary tolerance bands were selected. Using the research measures, these tolerance bands were changed to reflect performance standards of experienced instructor pilots. Using these tolerance bands based on expert performance, the percentage of time above, within, and below are computed. An overall score is also computed which reflects the percentage of time that all parameters are within tolerance simultaneously.

When the automated scoring is activated, these scores are presented following the completion of the maneuver. The results are displayed on a CRT at the instructor console. The instructor has the option to present the same information to the student by an in-cockpit CRT. To illustrate, the CRT displays are presented for the steep turn and takeoff in Figures 1 and 2.

A data storage and retrieval system, the Student Data System, was also developed to be used in conjunction with the operation of the automated performance measurement system. For each maneuver, the research measures are computed in addition to those instructional/feedback measures presented on the console CRT. In order to access these research measures, the Student Data System is used for initial storage and later retrieval. Analysis programs were developed for the processing of data directly from the student files, thereby eliminating manual data handling.

Future Developments

The most pressing requirement will be the completion of measurement packages for the remaining maneuvers taught in T-37 pilot training. Within the same time frame, it is expected that a series of evaluations will be accomplished on those maneuvers for which automated measurements are now available. It is anticipated that these evaluations will lead to a revision of the measurement scenarios. Once the measurement system is thoroughly evaluated and revised, a set of specifications will be written which should aid the development of performance measurement requirements for future simulation systems.

Table 1. Automatic Performance Measurement Maneuvers

A. BASIC CONTACT/INSTRUMENT
   - STRAIGHT AND LEVEL
   - TURNS TO HEADINGS
   - STEEP TURNS
   - CLIMBS/DESCENTS
   - SLOW FLIGHT
   - AIRSPEED INCREASE
   - AIRSPEED DECREASE
   - CONFIGURATION CHANGE

B. ADVANCED CONTACT
   - TAKEOFF
   - STRAIGHT-IN LANDING
   - OVERHEVD TRAFFIC PATTERN
   - OUTSIDE TRAFFIC PATTERN
   - AILERON ROLL
   - BARRELL ROLL
   - LOOP

4

C. ADVANCED INSTRUMENT
   - AUTOMATIC GCA
   - VERTICAL S A
   - VERTICAL S D
   - VOR COURSE INTERCEPT

7

	ALTITUDE	AIRSPEED	BANK	OVERALL SCORE
% HIGH	52.15	000000000	31.91	
% ON	47.84	63.44	63.82	27.02
% LOW	000000000	36.55	4.255	

INSTRUCTIONS:
1. SET POWER 81% : CLEAN
2. PERFORM 60 DEG BANK TURN, EITHER DIRECTION, AFTER COMMAND "ROLL IN"
3. MAINTAIN 15M', 160K
4. ROLL OUT ON COMMAND

Figure 1. Task 035, Steep Turn

```
                              OVERALL      47.10
                              SCORE


                          TAKE OFF     VERT VEL     CLIMB OUT
            HEADING       ATTITUDE     500-1000     ALTITUDE

% HIGH       27.27         78.78       000000000     8.064

% ON         72.72         18.18       100.0        79.03

% LOW        000000000      3.030      000000000    12.90

            MOVING AT BRAKE RELEASE        TRUE
            ROTATION AT:  72.22

             LATE- - - - - - - - - - -     TRUE

            LIFT OFF AT:  81.65

             UNSAFE- - - - - - - - - -     TRUE

            GEAR AT:      101.2

             UNSAFE AIRSPEED - - - - -     FALSE
             IN A DESCENT- - - - - - -     FALSE

            FLAPS AT:     113.5

             BEFORE GEAR                   FALSE
             BEFORE 110 KIAS               FALSE
             AFTER  135 KIAS - - - - -     FALSE

 WINDSPEED/DIRECTION   000000000 000000000  TURB  000000000
```

Figure 2. Task 105, Take Off Williams-30 Center

THE DECISION MEASUREMENT SYSTEM - CURRENT STATUS

F.B. Braun
M.A. Robinson
J.E. Tindall, Jr.

DATA-DESIGN LABORATORIES
Norfolk, Virginia 23502

THE DECISION MEASUREMENT SYSTEM - CURRENT STATUS

Introduction

The Decision Measurement System (DMS) is a multiple-choice, skill-related test that was developed by Data-Design Laboratories under the sponsorship of the Navy Strategic Systems Project Office.

The DMS represents an effort to develop a cost-effective alternative to "hands-on" performance testing of Fleet Ballistic Missile Weapons System (FBMWS) technicians. The DMS employs realistic simulations of equipment by presenting the panel displays, waveforms, diagnostic printouts, etc. that are associated with equipment operation or maintenance.

The DMS is being incorporated into the Personnel and Training Evaluation Program (PTEP) which is coordinated by the Central Test Site (CTS), Naval Guided Missiles School, Dam Neck, Virginia. PTEP is tasked with the assessment of the knowledge and skill levels of FBMWS personnel upon completion of formal training or while assigned to submarines, tenders, and training facilities.

Implementation Status

Each DMS consists of 12-20 exercises related to a piece of equipment or subsystem found on the FBM submarine. Each exercise, in turn, consists of 7-16 steps of an established (documented) operational or maintenance procedure. Each step in the DMS consists of a "Note" which poses a situation requiring a decision among four alternatives.

The DMS set consists of a Note Book and a Data Book. A sample page from one of the Note Books is shown in Figure 1.

Note N075 is an example of an initial note which in this case represents the starting point of a troubleshooting sequence. This note refers to two pictorials, P118 and P204, Figures 2 and 3, respectively, and a diagnostic printout T118, Figure 4.

The examinee indicates his decisions by marking an Exercise Sheet (answer form), Figure 5, with a special pen. This results in the appearance of a code number (Figure 5, Block 1) which has been printed on the exercise sheet with invisible ink. The code number directs him to the next step in the sequence (N167) if he made the right decision (A) or it directs him to pick another alternative if he made the wrong decision (B, C, or D). The examinee will invariably progress to the next step in the sequence.

Thus, the DMS presents a simulated situation, Figure 6, in which a problem is stated, related pictorial and/or symbolic data is presented, and a decision is required. The score is based upon the number of decisions that are made by the examinee in order to complete a given sequence. This scoring scheme was described in detail at last year's meeting. It should be noted that the answer form has mark-sense areas next to each answer that are filled in by the examinee after he has completed the test; these provide a means for machine reading and automatic scoring of the DMS.

We are in the process of developing approximately 25 DMS sets for use in PTEP. To date, eight of these tests have been delivered to CTS. Only one set has been in use long enough to obtain statistical data. This is the DMS for the Missile Test and Readiness Equipment, Mk 7, Mod 2, generally referred to as the MTRE-7. This test has been administered to over 200 Missile Technicians and forms the basis for the preliminary data reported here.

Developmental Problems

As we had anticipated, a major problem associated with the development of a DMS was obtaining qualified technical personnel and training them to develop test items in the DMS framework.

To obtain technical experts, we have generally sought present or past instructors of the specific equipment or subsystem to be covered in a DMS, and have provided them with intensive training in DMS construction techniques. We have found that it takes at least three weeks to train a technical expert to develop a DMS--one week of tutoring (we usually train one person at a time) and two weeks of close monitoring.

The most difficult aspect of this training is to develop a sense of judgment as to the proper level of difficulty of the test items and the selection of sufficient specific data (panel displays, waveforms, etc.) to provide for a definitive solution to the problem presented in each note.

Another problem that we have encountered is the availability of equipment to verify the DMS. This equipment is available to us only at the training sites, and requires considerable coordination of schedules to assure sufficient time on a non-interference basis with training activities.

Test Results

We are planning to obtain data on the actual performance of FBMWS technicians and correlate these performance data with scores obtained on selected DMS tests. Since FBM equipment seldom requires repair during normal patrols, we are exploring the use of "prefaulted" module tests which can be administered at the training sites and which will provide the required performance data.

In the interim, we have developed a statistical indication of test validity, arrived at by separating out those technicians who were directly involved with MTRE-7 maintenance (n = 20) from the population of FBM Missile Technicians (n = 210) who took the DMS. The average score obtained on the DMS by MTRE-7 technicians was significantly greater than the average score of the Missile Technicians who were not directly involved with the MTRE-7, as shown in Table 1. It should be noted that the aptitude scores and educational background of the MTRE-7 technicians and non-MTRE-7 Missile Technicians showed them to be almost identical in these respects.

As noted above, PTEP measures both the knowledge and skill of FBMWS technicians with respect to specific equipment or subsystems associated with their respective specialties. In the case of Missile Technicians, there is a knowledge area which deals with the MTRE-7 equipment. This made it possible to determine what differences, if any, existed in this specific knowledge area between MTRE-7 technicians and other (non-MTRE-7) Missile Technicians.

Table 2 gives the specific MTRE-7 knowledge scores for MTRE-7 technicians versus other Missile Technicians. The numbers differ somewhat from Table 1 since not all Missile Technicians took the knowledge test. It can be seen that although the MTRE-7 technicians did better than the other Missile Technicians, the difference was not significant. This is an interesting finding when contrasted with the scores on the DMS (Table 1). It suggests that the DMS is a better indicator of hands-on experience than the MTRE-7 knowledge test. We intend to explore this hypothesis further as DMS test data are accumulated.

Table 3 gives the correlations (Pearson r) between the DMS and MTRE-7 knowledge scores for Missile Technicians. The highest correlation (.53) was obtained for the MTRE-7 technicians. The significance of this correlation is at the .02 level, based upon the number of individuals in the group. The next highest correlation (.38) was obtained for E5s (Enlisted Grade 5). This correlation is significant at the .01 level, as is the correlation for Missile Technicians as a whole (.31). It is interesting to note that the correlation for grades E7/E8 is near zero. Inspection of the actual scores for these two grades of Missile Technicians (which are largely supervisory) showed that on the whole, they did better than the other grades in MTRE-7 knowledge, and just about as well as the other grades on the DMS. However, the low correlation indicates that there was not any significant "coupling" of the scores on the two tests, and that may be the result of lack of recent hands-on experience in these two grades.

Thus, we have an additional hypothesis to be tested; namely, that hands-on experience leads to the integration of knowledge and skill, resulting in higher correlations between scores in these areas.

Future Plans

The results reported here are based upon a single DMS administered to a group of approximately 200 Missile Technicians. However, we feel that these findings will be sustained with larger numbers of examinees.

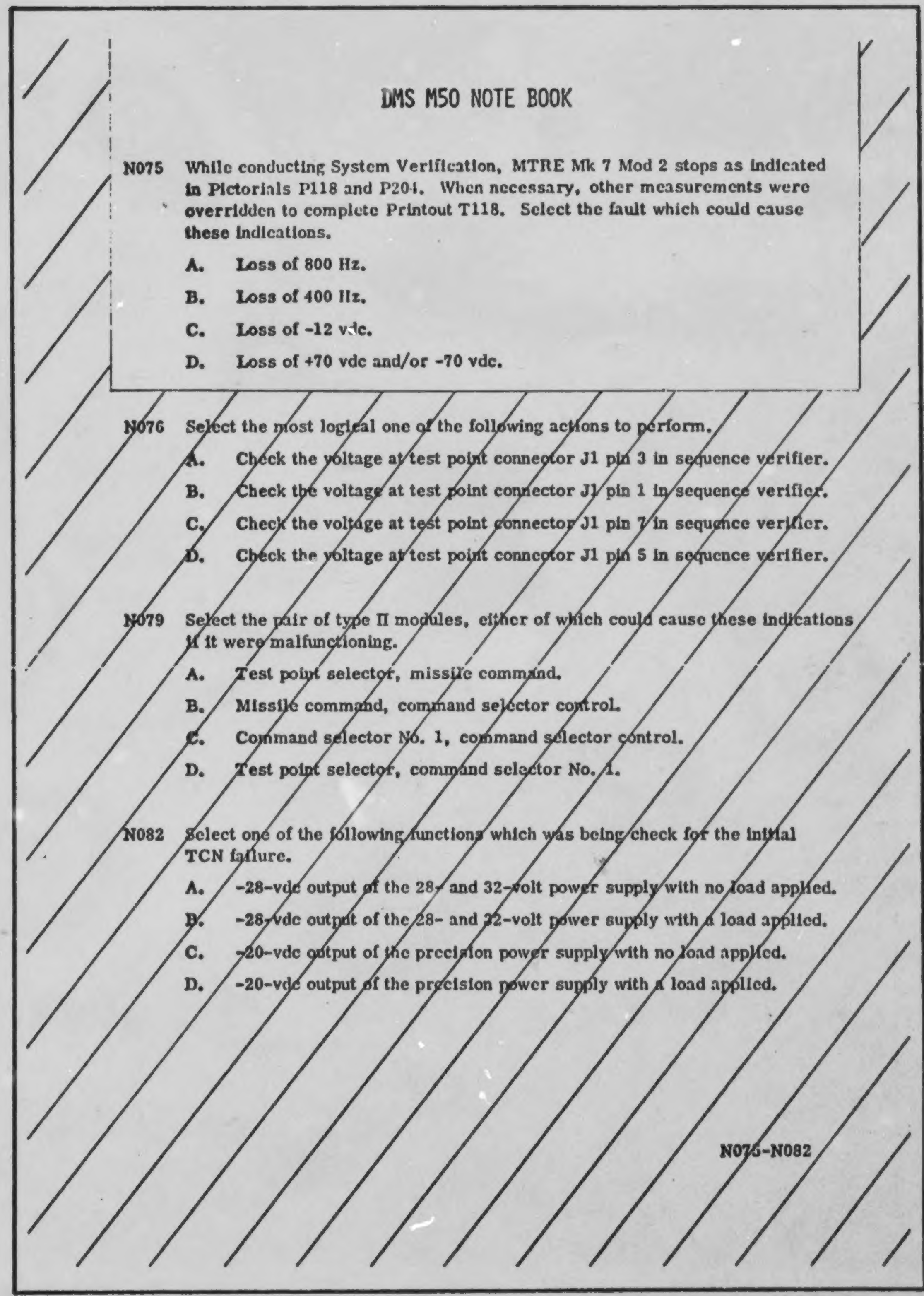
We have begun to study some of the "structural" factors of the DMS; i.e., the specific classes of skills associated with the Notes (test items), the types of recurrent errors made, and the relationship among various parts of the DMS.

We are planning to develop several alternative scoring schemes, based upon these structural factors, as well as on the time required to complete the DMS, and will compare these scoring schemes with the one we presently use.

It should be noted that the Navy Personnel Research and Development Center (NPRDC) in San Diego, California is planning to use the MTRE-7 DMS in connection with a personnel training research project involving correlation of various tests with shipboard performance. These data should be made available to us in the near future.

SUMMARY

We believe that we have developed a unique skill-related test that is applicable to a wide variety of situations in which tasks are sequentially organized. By using latent-image techniques, we are able to direct an individual down a given solution path and, in the process, assess his ability to pursue a logical course of action based upon the information he receives at each step in the sequence. We are planning to conduct a variety of validation studies of this test as well as to examine its internal structure more comprehensively.



DMS M50 NOTE BOOK

N075 While conducting System Verification, MTRE Mk 7 Mod 2 stops as indicated in Pictorials P118 and P204. When necessary, other measurements were overridden to complete Printout T118. Select the fault which could cause these indications.

- A. Loss of 800 Hz.
- B. Loss of 400 Hz.
- C. Loss of -12 vdc.
- D. Loss of +70 vdc and/or -70 vdc.

N076 Select the most logical one of the following actions to perform.

- A. Check the voltage at test point connector J1 pin 3 in sequence verifier.
- B. Check the voltage at test point connector J1 pin 1 in sequence verifier.
- C. Check the voltage at test point connector J1 pin 7 in sequence verifier.
- D. Check the voltage at test point connector J1 pin 5 in sequence verifier.

N079 Select the pair of type II modules, either of which could cause these indications if it were malfunctioning.

- A. Test point selector, missile command.
- B. Missile command, command selector control.
- C. Command selector No. 1, command selector control.
- D. Test point selector, command selector No. 1.

N082 Select one of the following functions which was being check for the initial TCN failure.

- A. -28-vdc output of the 28- and 32-volt power supply with no load applied.
- B. -28-vdc output of the 28- and 32-volt power supply with a load applied.
- C. -20-vdc output of the precision power supply with no load applied.
- D. -20-vdc output of the precision power supply with a load applied.

N075-N082

Figure 1.

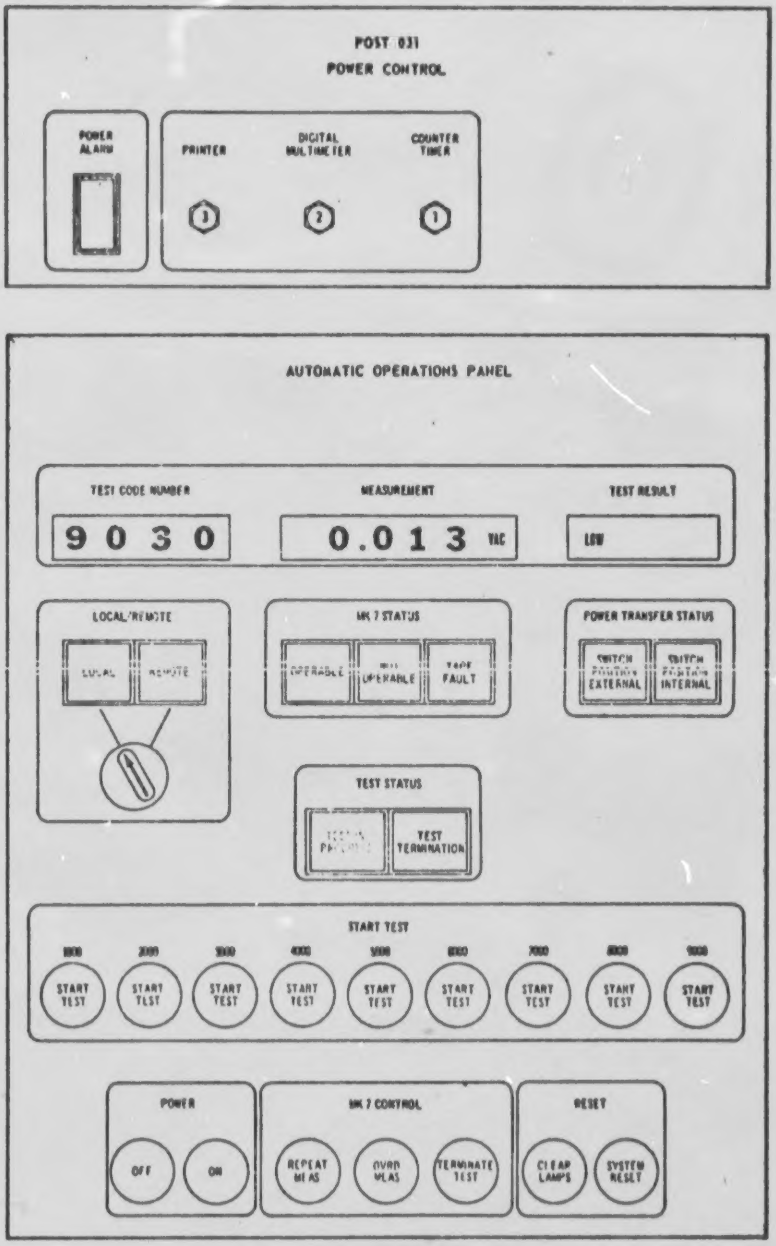
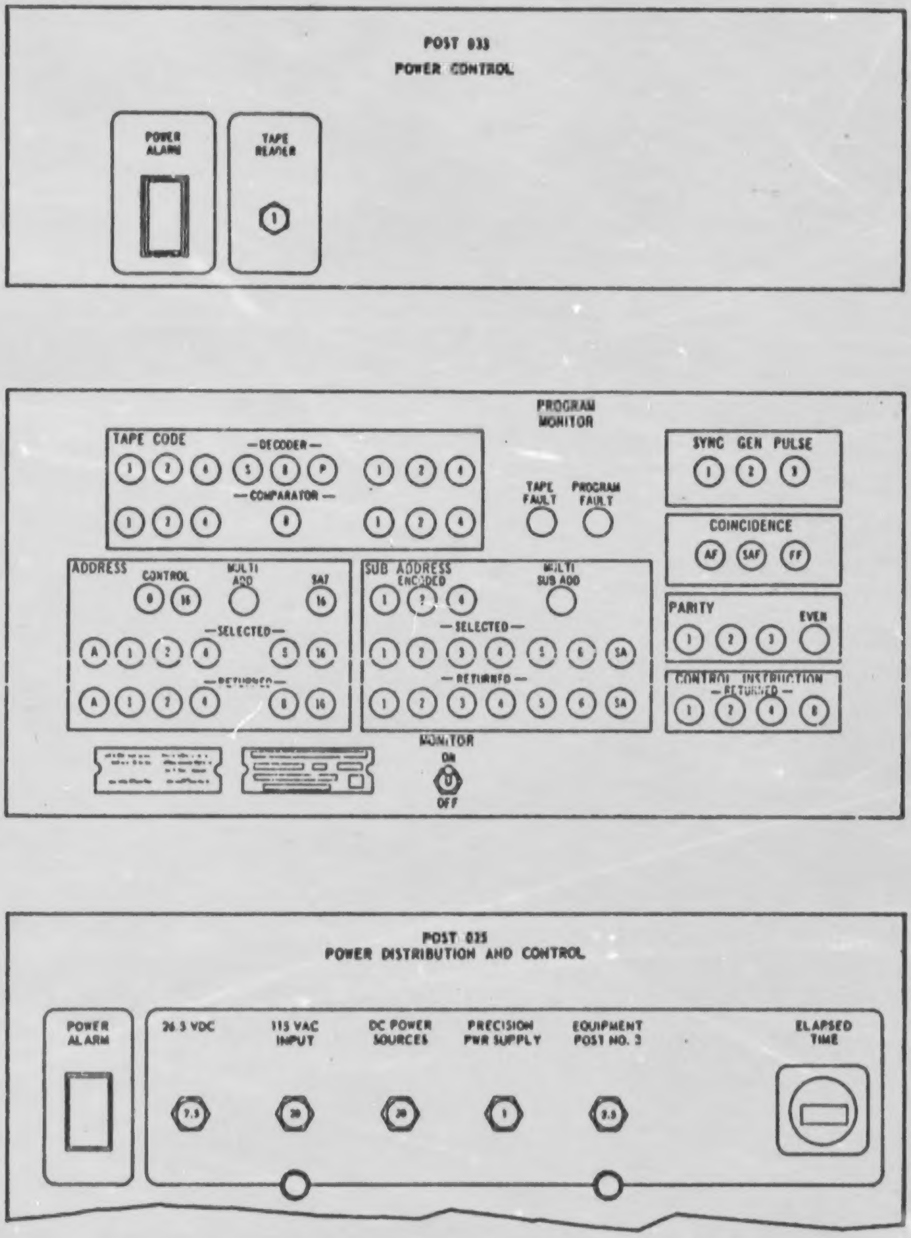

POST 031
POWER CONTROL
POWER ALARM
PRINTER
DIGITAL MULTIMETER
COUNTER TIMER
3
2
1
AUTOMATIC OPERATIONS PANEL
TEST CODE NUMBER
MEASUREMENT
TEST RESULT
9 0 3 0
0.0 1 3 VAC
LOW
LOCAL/REMOTE
LOCAL
REMOTE
MK 7 STATUS
OPERABLE
INOP OPERABLE
TAPE FAULT
POWER TRANSFER STATUS
SWITCH POSITION EXTERNAL
SWITCH POSITION INTERNAL
TEST STATUS
TEST IN PROGRESS
TEST TERMINATION
START TEST
1000
2000
3000
4000
5000
6000
7000
8000
9000
START TEST
POWER
OFF
ON
MK 7 CONTROL
REPEAT MEAS
OVRD MEAS
TERMINATE TEST
RESET
CLEAR LAMPS
SYSTEM RESET

Figure 2.


POST 033
POWER CONTROL
POWER ALARM
TAPE READER
1
PROGRAM MONITOR
TAPE CODE
— DECODER —
1 2 4 S 8 P 1 2 4
— COMPARATOR —
1 2 4 8 1 2 4
TAPE FAULT
PROGRAM FAULT
SYNC GEN PULSE
1 2 3
COINCIDENCE
AF SAF FF
ADDRESS
CONTROL
8 16
MULTI ADD
SAF
16
— SELECTED —
A 1 2 4 8 16
— RETURNED —
A 1 2 4 8 16
SUB ADDRESS
ENCODED
1 2 4
MULTI SUB ADD
— SELECTED —
1 2 3 4 5 6 SA
— RETURNED —
1 2 3 4 5 6 SA
PARITY
1 2 3
EVEN
CONTROL INSTRUCTION
— RETURNED —
1 2 4 8
MONITOR
ON
OFF
POST 035
POWER DISTRIBUTION AND CONTROL
POWER ALARM
26.5 VDC
7.5
115 VAC INPUT
20
DC POWER SOURCES
20
PRECISION PWR SUPPLY
1
EQUIPMENT POST NO. 3
2.5
ELAPSED TIME

P204

Figure 3.

DMS M50 DATA BOOK

9302	9999	3	S	GO
9301	9999	3	S	GO
9300	9999	3	S	GO
9262	2476	2	AC	GO
9261	2474	2	AC	GO
9260	2476	2	AC	GO
9258	6494	3	AC	GO
9257	9502	3	AC	GO
9256	2587	2	AC	GO
9255	2985	2	AC	GO
9254	3288	2	AC	GO
9253	2282	2	AC	GO
9252	1583	2	AC	GO
9251	0004	4	AC	GO
9250	8000	3	AC	GO
9247	0016	2	AC	GO
9246	2376	2	AC	GO
9241	0017	3	AC	GO
9240	7485	3	AC	GO
9232	1580	2	AC	GO
9231	0000	4	AC	GO
9230	8000	3	AC	GO
9227	0017	2	AC	GO
9226	2373	2	AC	GO
9221	0017	3	AC	GO
9220	7486	3	AC	GO
9218	6493	3	AC	GO
9217	9501	3	AC	GO
9216	2585	2	AC	GO
9215	2986	2	AC	GO
9214	3287	2	AC	GO
9213	2284	2	AC	GO
9212	1583	2	AC	GO
9211	0009	4	AC	GO
9210	7994	3	AC	GO
9207	0017	2	AC	GO
9206	2378	2	AC	GO
9201	0018	3	AC	GO
9200	7486	3	AC	GO
9124	0030	4	+	GO
9123	9999	4	-	LO
9122	0245	4	+	GO
9121	0033	4	+	GO
9120	6998	3	+	GO
9119	5730	3	+	LO
9118	6972	3	+	GO
9117	6992	3	+	GO
9116	9495	3	+	GO
9115	5790	3	+	LO
9114	9478	3	+	GO
9113	9505	3	+	GO

T117 (Cont)

9008	1200	2	-	GO
9007	6001	3	+	GO
9006	1200	2	-	GO
9005	5970	3	+	GO
9004	3205	2	+	GO
9003	2795	2	-	GO
9002	2800	2	+	GO
9001	2780	2	+	GO
0041	0000	1	KC	GO
0688	0000	1	KC	GO
0014	0000	1	KC	GO
0072	0000	1	KC	GO

T118

9861	0000	2	-	GO
9860	7996	3	-	LO
9842	1135	2	-	GO
9841	0027	2	+	GO
9832	0000	2	-	GO
9831	2993	2	+	GO
9822	8612	3	-	GO
9821	3997	3	-	GO
9820	2001	2	-	GO
9812	0000	2	-	GO
9811	3207	2	+	GO
9810	3204	2	-	GO
9520	0232	3	S	GO
9343	1151	2	S	GO
9342	4472	3	S	GO
9341	0437	3	S	GO
9340	0040	3	S	GO
9333	3721	3	S	GO
9332	2784	3	S	GO
9331	1847	3	S	GO
9330	0910	3	S	GO
9327	0225	2	S	GO
9326	0225	2	S	GO
9325	0225	2	S	GO
9324	0225	2	S	GO
9323	6695	3	S	GO
9322	6695	3	S	GO
9321	6695	3	S	GO
9320	6695	3	S	GO
9313	0066	3	S	GO
9312	0066	3	S	GO
9311	0066	3	S	GO
9310	0066	3	S	GO
9303	9999	3	S	GO

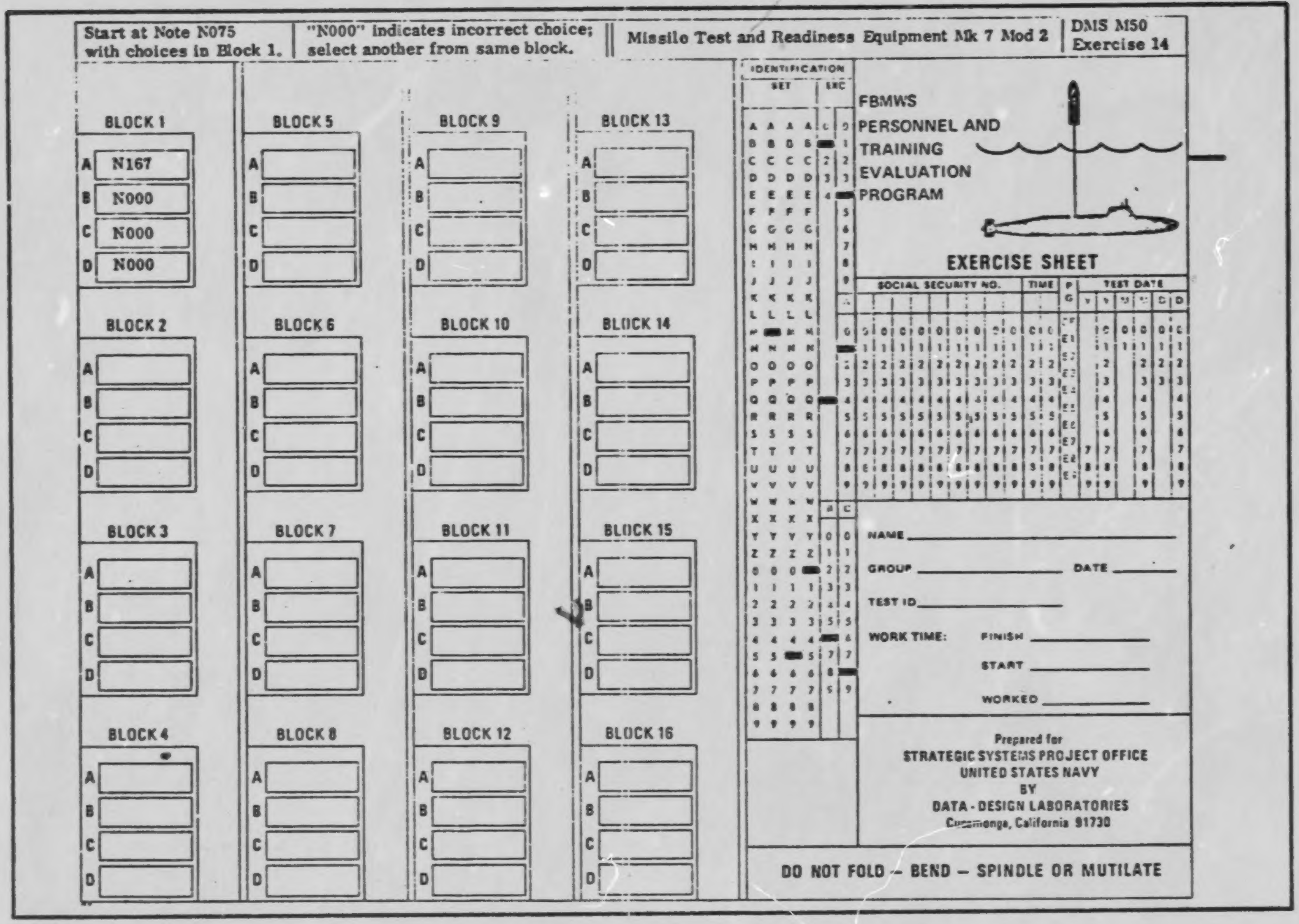
T117 (Cont)

9107	9208	3	+	HI
9106	0387	3	+	HI
9105	9999	3	-	LO
9104	9999	3	-	LO
9103	9999	3	-	LO
9102	1838	3	-	HI
9101	0000	3	-	HI
9073	0000	2	+	LO
9072	0000	2	-	HI
9071	2002	2	-	GO
9093	9999	4	-	LO
9092	6440	3	+	HI
9091	0601	3	-	HI
9090	9999	4	-	LO
9089	9999	4	-	LO
9088	9999	4	-	LO
9087	6623	3	+	HI
9086	5027	3	+	HI
9085	9999	4	-	LO
9084	9999	4	-	LO
9083	9999	4	+	HI
9082	5985	3	-	HI
9081	9999	3	+	HI
9080	9999	4	-	LO
9079	9999	4	-	LO
9078	9999	4	-	LO
9077	9999	3	-	LO
9076	9999	3	+	HI
9075	9999	4	-	LO
9074	9999	4	-	LO
9073	0001	2	+	LO
9072	0000	2	-	HI
9071	2002	2	-	GO
9061	0000	0	CY	LO
9060	0000	0	CY	LO
9051	0000	0	CY	LO
9043	9999	1	KΩ	GO
9042	9531	2	KΩ	GO
9041	1004	3	KΩ	GO
9040	0021	4	KΩ	GO
9031	0018	2	AC	LO
9030	0013	3	AC	LO
9026	6346	2	+	GO
9025	3197	2	-	GO
9024	1600	2	+	GO
9023	7998	3	-	GO
9022	8991	3	+	GO
9021	0999	3	+	GO
9020	5007	4	+	GO
9010	1200	2	-	GO
9009	5995	3	+	GO

T118 (Cont)

T117-T118

Figure 4.



Start at Note N075 with choices in Block 1.

"N000" indicates incorrect choice; select another from same block.

Missile Test and Readiness Equipment Mk 7 Mod 2

DMS M50
Exercise 14

BLOCK 1

A N167
B N000
C N000
D N000

BLOCK 2 · BLOCK 3 · BLOCK 4 · BLOCK 5 · BLOCK 6 · BLOCK 7 · BLOCK 8 · BLOCK 9 · BLOCK 10 · BLOCK 11 · BLOCK 12 · BLOCK 13 · BLOCK 14 · BLOCK 15 · BLOCK 16

IDENTIFICATION
SET
EXC

FBMWS
PERSONNEL AND
TRAINING
EVALUATION
PROGRAM

EXERCISE SHEET

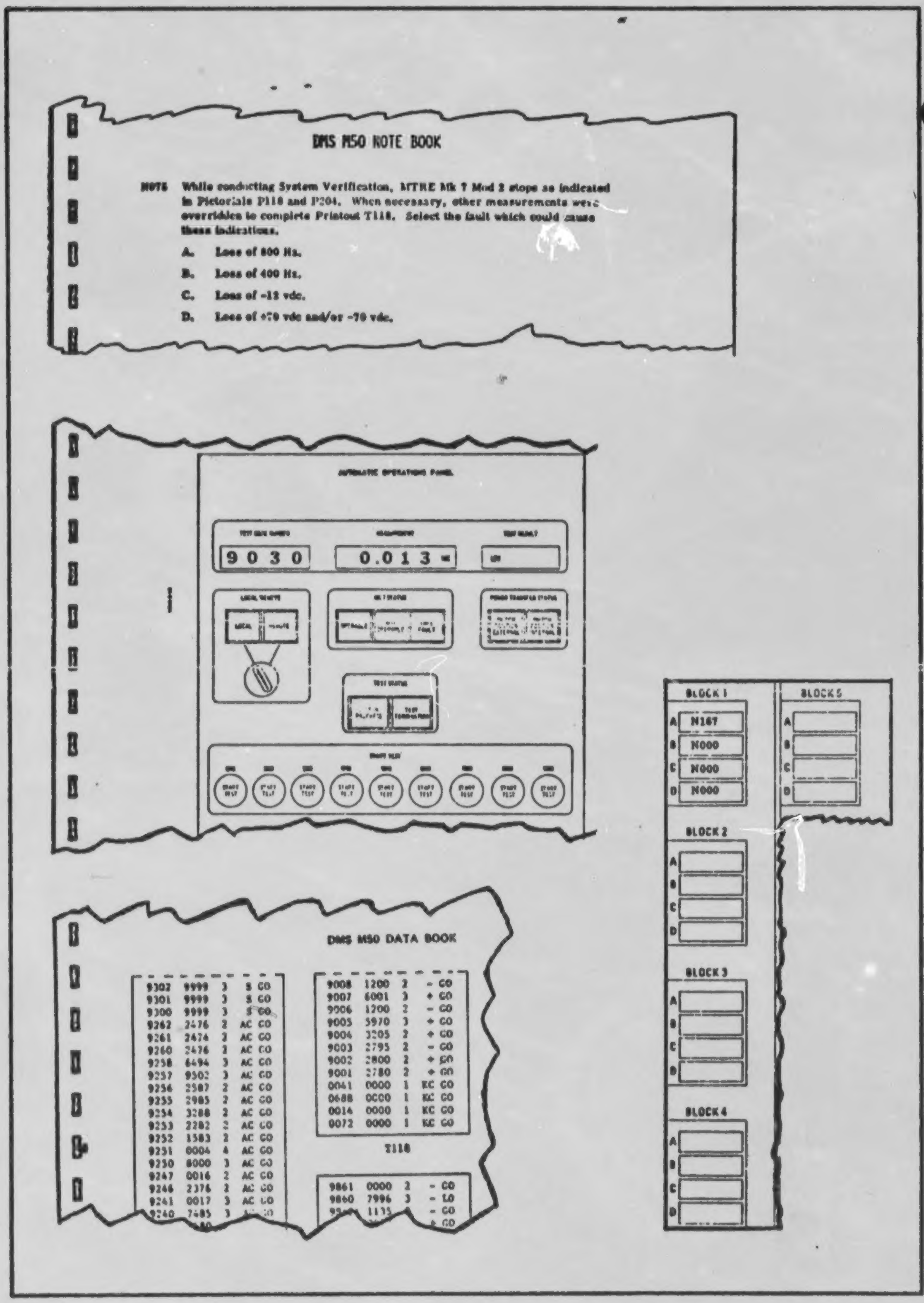
SOCIAL SECURITY NO. | TIME | P G | TEST DATE

NAME ____
GROUP ____ DATE ____
TEST ID ____
WORK TIME: FINISH ____
START ____
WORKED ____

Prepared for
STRATEGIC SYSTEMS PROJECT OFFICE
UNITED STATES NAVY
BY
DATA - DESIGN LABORATORIES
Cucamonga, California 91730

DO NOT FOLD — BEND — SPINDLE OR MUTILATE

Figure 5.



DMS M50 NOTE BOOK

N075 While conducting System Verification, MTRE Mk 7 Mod 2 stops as indicated in Pictorials P118 and P204. When necessary, other measurements were overridden to complete Printout T118. Select the fault which could cause these indications.

A. Loss of 800 Hz.

B. Loss of 400 Hz.

C. Loss of -12 vdc.

D. Loss of +70 vdc and/or -70 vdc.

AUTOMATIC OPERATIONS PANEL

DMS M50 DATA BOOK

```
9302   9999   3    S GO
9301   9999   3    S GO
9300   9999   3    S GO
9262   2476   2   AC GO
9261   2474   3   AC GO
9260   2476   3   AC GO
9258   6494   3   AC GO
9257   9502   3   AC GO
9256   2587   2   AC GO
9255   2985   2   AC GO
9254   3288   2   AC GO
9253   2282   2   AC GO
9252   1583   2   AC GO
9251   0004   4   AC GO
9250   8000   3   AC GO
9247   0016   2   AC GO
9246   2376   2   AC GO
9241   0017   3   AC LO
9240   7485   3      GO
```

```
9008   1200   2    - GO
9007   6001   3    + GO
9006   1200   2    - GO
9005   5970   3    + GO
9004   3205   2    + GO
9003   2795   2    - GO
9002   2800   2    + GO
9001   2780   2    + GO
0041   0000   1   KC GO
0688   0000   1   KC GO
0014   0000   1   KC GO
0072   0000   1   KC GO
```

T118

```
9861   0000   2    - GO
9860   7996   3    - LO
       1135        - GO
                   + GO
```

Figure 6.

TABLE 1

DMS SCORES FOR MISSILE TECHNICIANS WITH AND WITHOUT MTRE-7 EXPERIENCE

GROUP	NUMBER	MEAN	S.D.
WITH MTRE-7 EXPERIENCE	20	238.9	35.8
WITHOUT MTRE-7 EXPERIENCE	190	217.2	47.4
		Difference = 21.7*	

*Difference between means is significant at the .01 level for a one tailed test (z = 2.49).

TABLE 2

MTRE-7 KNOWLEDGE SCORES FOR MISSILE TECHNICIANS WITH AND WITHOUT MTRE-7 EXPERIENCE

GROUP	NUMBER	MEAN	S.D.
WITH MTRE-7 EXPERIENCE	18	17.4	3.3
WITHOUT MTRE-7 EXPERIENCE	84	16.3	3.9
Difference = 1.1*			

*Difference between means not significant ($z = 1.24$)

TABLE 3

CORRELATIONS BETWEEN DMS AND MTRE-7 KNOWLEDGE SCORES FOR MISSILE TECHNICIANS

GROUP	NUMBER	PEARSON r	SIGNIFICANCE
E3/E4	46	.20	NS*
E5	99	.38	.01
E6	29	.30	NS
E7/E8	28	-.05	NS
ALL	202	.31	.01
MTRE-7 TECH's		.53	.02

*NS - not significant

TASK ANALYSIS:

THE BASIS FOR CRITERION TEST AND CURRICULUM DESIGN

By

Robert N. Johnson and 1LT James N. Richmann
U.S. Army Institute of Administration
Fort Benjamin Harrison, IN 46216

Recently the civilian educational community has become enamored with the concept of educational accountability, a philosophy that we in the military have always followed. Simply stated, educational accountability requires specification of the desired results of the system, measurement of the instruction in terms of these results, and a continued search for ways to lower costs or improve the program. In the military the foundation for attaining accountability is task analysis, the process of specification of training outcomes in terms of job actions, job conditions, and results of job performance. Task analysis is the basis for development of the criterion tests which will be used to measure instruction. Task analysis is also the basis for the design of the curriculum, and therefore the cost and quality of this curriculum is tied directly to the quality of that task analysis. But what is the quality of our current task analyses? Do they accomplish the purposes for which they are intended? Dr. Charles Jackson of the US Army Armor School in the report of an Instructional Technology Study Group observed that "personnel performing task analysis activities were preparing task lists of imprecise tasks and were not analyzing these tasks in sufficient detail to facilitate instructional design." This paper addresses these task analysis problems, proposes solutions and suggests a methodology for the use of a well documented task analysis in criterion test and curriculum design.

In addressing the problems inherent in task analysis it is important to identify a conceptual model for the efficient use of task analysis information in the instructional design process. Such models are common and almost all agree as to what task analysis should accomplish. A typical model, simplified to chart the relationship of task analysis to other instructional design products is shown on this transparency (See Figure 1)

INSTRUCTIONAL DESIGN PRODUCTS

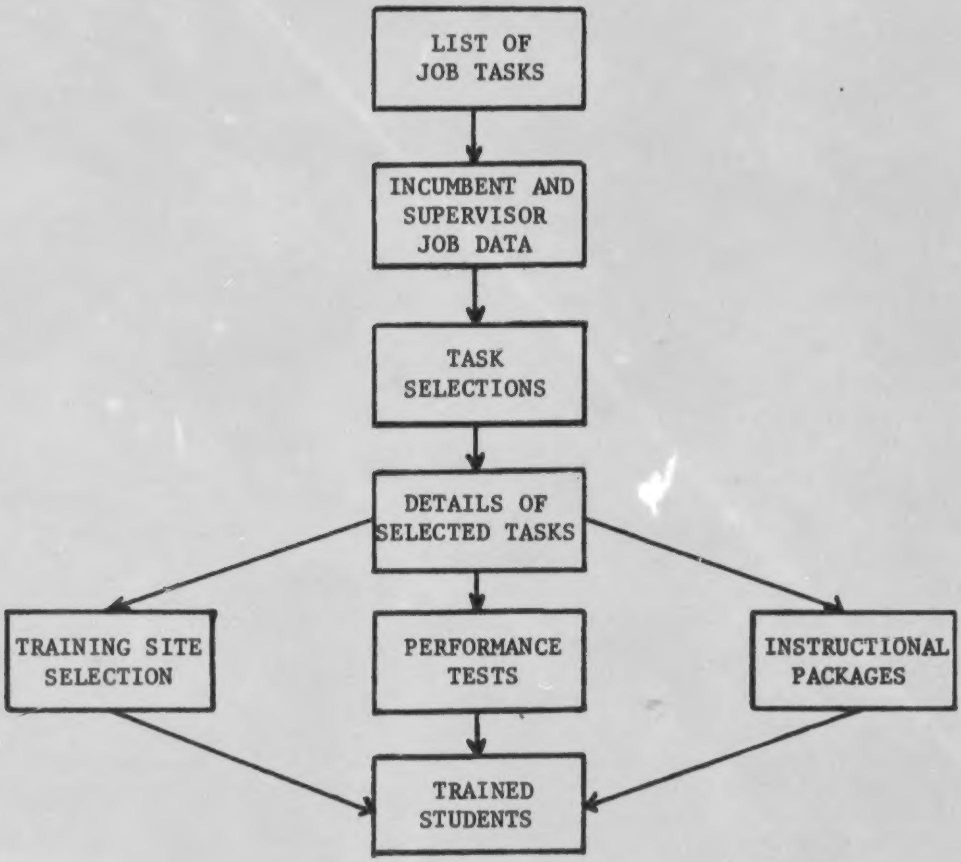

LIST OF JOB TASKS
INCUMBENT AND SUPERVISOR JOB DATA
TASK SELECTIONS
DETAILS OF SELECTED TASKS
TRAINING SITE SELECTION
PERFORMANCE TESTS
INSTRUCTIONAL PACKAGES
TRAINED STUDENTS

Figure 1

The first requirement is to make a list of tasks performed by job incumbents. Second, we need to collect job data from job incumbents and supervisors which will permit a cost-effective selection of tasks for training and testing. Third, we need to analyze each task into the detail required for selection of training site, design of performance tests, and development of instructional packages or lessons.

Examination of these products indicates that two levels of task specificity are required in instructional design (See Figure 2).

LEVELS OF TASK DETAIL

DETAIL	USED FOR
Short Task Statement	Task Inventory Data Collection Task Selection
Detailed Task Description	Training Site Selection Test Design Instructional Package Design

Figure 2

First, short terse statements of job tasks are used to make the task inventory and to collect the data necessary for task selection. Task statements which are long or detailed do not facilitate data collection and result in a cumbersome task list. Second, detailed descriptions of each task are used to select the training sites, design the performance tests, and develop the instructional packages. The terse task statements used to collect job data do not contain sufficient detail for these later functions. They may tell you what is done, but do not tell you how. To select a training site, design a performance test, or develop an instructional package, the designer needs to know the cues to performance, the elements of performance, the conditions which affect performance, and the results of acceptable task performance.

Instructional designers have recognized these two different levels of detail, but have not recognized the relationship between the two.

With respect to the need for short terse task statements, the major problem has been levels of specificity. Task inventories have been prepared which contain statements which are often too broad to permit further detailed analysis or are too narrow to be considered independent tasks. As a result, data is collected and selections are made only to find out that the original task statements are inadequate to serve as the basis for future steps in instructional design

For examples of this problem of level of specificity, consider these examples (See Figure 3)extracted from the task inventory for the job of unit First Sergeant.

EXTRACT OF TASK LIST

JOB = COMPANY FIRST SERGEANT

SUPERVISE UNIT ADMINISTRATION SAFEGUARD CLASSIFIED MATERIALS	TOO BROAD
SELECT A DETAIL USING A DUTY ROSTER INVENTORY CLASSIFIED DOCUMENTS	ACCEPTABLE
DETERMINE CAUSE OF SOLDIER'S ABSENCE LABEL CLASSIFIED MATERIAL	TOO NARROW

Figure 3

Some task statements are so broad as to preclude detailed analysis since the exact parameters of the task are not defined. Others are so narrow that analysis reveals they are really elements of a larger task. Consider the first two task statements on the list. These tasks have a different meaning and interpretation for each of us. Each possible interpretation would result in a different detailed analysis of the task and thus change both the training and test content. If we do not catch this type of task prior to the collection of job data, then problems will arise in the selection of tasks for training. Varying levels of specificity distort values of "Relative time spent" and other statistics calculated from cumulative job data. Even the raw data for this task will be suspect since no one can be certain which portion of the responses pertain to the various interpretations of the task. In effect, we are not in any better position to design training than we were when we used subjects rather than tasks as the basis for instructional design.

Now, consider the last two task statements on the list. A cursory examination of these tasks shows no real problems, but a detailed analysis reveals that these tasks are really part of a larger, more complete and cohesive, task. It makes little sense to collect job data on these elements of performance when detailed analysis of the larger task may establish that total task can not be adequately performed without that element. In addition, including task elements like these in the inventory increases the administrative burden of the data collection.

Our objective is to write task statements at the level of specificity which represents a meaningful "whole" on the job from the viewpoint of the job incumbent. The only solution to this problem of task specificity is a careful edit of the task inventory prior to data collection, combined with proper use of iterative feedback from the further detailed task analysis. Management must realize that no design product, including a task inventory, is initially perfect. The "trial-test-revise" procedures of validation of instructional materials are applicable during the process of task analysis as well. If, during the detailed analysis of a task statement we find the original level of specificity was too broad or too narrow, then the task inventory must be changed.

In editing the task list prior to the collection of job data, the analyst must carefully consider each statement and determine, not only whether it is a task, but also whether or not it is written at a level of specificity which facilitates further detailed analysis. Although, we utilize the commonly accepted definitions of a task, we have added a new dimension. In addition to the requirements that a task is a highly specific unit of job activity stated in job terms with an action verb and an object, must have a definite beginning and a definite end, must have at least one cue or stimulus which, in the real world signals the incumbent to perform that task and must be observable and measurable, we insist that a task must be the smallest unit of job activity performed for its own sake in the eyes of a job incumbent in the job situation. (See Figure 4)

A TASK MUST BE THE SMALLEST UNIT OF JOB ACTIVITY DONE FOR ITS OWN SAKE IN THE EYES OF THE JOB INCUMBENT IN THE JOB SITUATION.

Figure 4

It is the smallest unit of meaningful performance to the Soldier. This is the key distinguishing characteristic between a task and an element of a task. For example, a mechanic does not remove the wheel and tire from a vehicle for the purpose of removing the wheel and tire. The removal is made in order to accomplish one of several tasks, e.g., "rotate the tires," "repair a flat," "install new tires," etc. These latter statements are tasks because they are done for their own sake in a job situation. Each of them requires, as an element of performance, the removal of the wheel and tire from the vehicle. By including "done for its own sake in a job situation" as the major characteristic of a task as opposed to an element, we permit differentiation among tasks and elements. By including "in the eyes of the job incumbent" we recognize that a statement which is a task at one job level may be a duty or only an element at another job level. By including "the smallest unit of meaningful performance" we recognize that duties and jobs, etc., are meaningful units of performance, but do not furnish sufficient detail to permit further analysis. While judgement of this criteria for a listed statement is subjective, the use of Robert Mager's "Hey-Dad" test* is a fairly good indicator of whether or not the statement is a task or an element. In this test one imagines asking the job incumbent "What are you doing now?" If the job incumbent could be expected to answer with the listed statement, then the statement is probably a task.

The preceding discussion emphasized the first requirement for specificity in task analysis--the need for short, terse task statements which represent meaningful "wholes" to job incumbents. Now, we would like to shift the discussion to the second requirement--the need for detailed task descriptions which provide the basis for design of tests and instructional packages. In relationship to the overall model of instructional design (See Figure 1) the production of these detailed task descriptions follows job data collection and task selection. The position of the detailed analysis in the process is an important factor in establishing a logical and cost effective sequence to the overall instructional design system. It is logical to perform the detailed analysis at this point since the information derived is not required earlier in the process. It is cost effective since it precludes detailed analysis of tasks which may be eliminated during the task selection procedure. The detailed task analysis is still at this stage, divorced from training. This is as it should be since we are concerned with how the task is performed in the real world, not how it will be performed during training. One of the gravest errors we could make during design would be to base our detailed analysis, and hence our tests, on the content of the training rather than on the actual job requirements. For this reason, we must separate the detailed task analysis from the training analysis which results in design of lessons, instructional packages, or media materials.

---

* Mager, Robert F. Goal Analysis, Belmont, CA: Fearon Publishers, 1972, page 30.

The central problem in developing a detailed task analysis is to find a way to document the work of the analyst. While forms, procedures, and the like are not ends in themselves, the most brilliant thought, if not carefully recorded for future reference, is lost and useless. Good documentation should be flexible, easy to understand, easy to reference, and have no unnecessary components. Since documentation is only a tool, it should be easy and cheap to construct and reproduce. Good documentation should direct the thinking of a mediocre analyst to discover the critical aspects of the task, but not restrict the thinking of the good analyst to discover imaginative new solutions to old problems. Because tasks are different, format should not dictate the analysis procedure. To a large extent, therefore, the usefulness of the detailed task analysis, and hence the quality of the following instructional design products is dependent on the type of documentation used.

In the U.S. Army Institute of Administration we have had considerable success documenting the detailed task analysis with a format we call a Task Structure Analysis to differentiate it from activities concerned with the task statement only. On the Task Structure Analysis form we document four major aspects of the task performance (See Figure 5).

TASK STRUCTURE ANALYSIS

DOCUMENTATION OF:

1. CUES

2. CONDITIONS

3. ELEMENTS OF PERFORMANCE

4. RESULTS

Figure 5

First, we record the cues which initiate the task. Second, we state the elements of performance in their logical sequence. The elements of performance are either the steps in the execution of the task, or the decisions made during performance which alter the sequence of the process. Next, in relation to each of the elements of performance we document the conditions which are associated with performance. These conditions may be tools, references, job aids, environmental or attitudinal factors which affect performance. Finally, we record the characteristics of the results of adequate task performance. These results of task performance may be documented either in terms of characteristics of an acceptable product, key steps in an acceptable process, or in terms of known errors in execution of the process.

At this time I would like to pass out a completed Task Structure Analysis to give you an example of the format and procedures used to record the detailed analysis (Pass Out Figure 6). The task documented on this example is "Select a Detail Using a Duty Roster." The cues which initiate task performance are first listed in block 7 on the form. You will notice that each cue is associated with a response consisting of execution of one of the elements of task performance. Most frequently, all cues are associated with the first element of performance, but this is not always the case. Some cues, for example, may complicate the situation and require extra steps or decisions prior to performance of the "main stream" elements. One of the major advantages of this form is its flexibility in showing this type of relationship among the cues.

Once at least one cue has been identified the designer may begin to list the elements of performance in block 8. Those of you who are familiar with decision flow charts or logic trees will recognize that this format is nothing more than a flow chart in narrative form. Instead of using boxes and arrows we use the associated "decision-yes-no" and "go-to-step" blocks to establish the relationships among the elements of performance. Again, the format is designed to maximize the flexibility for the designer. You may use this form to document a stimulus-response table, a linear list of sub-tasks, or a decision table. In our experience, however, most real world tasks have considerable branching in that varying conditions and cues affect the manner in which the task is performed. Sorting out the logic of these interactions is a time consuming process, but one which yields considerable insight. Seldom is a simple list of subtasks sufficient to detail the complexity of real world task performance. The Task Structure Analysis form forces the designer to consider the interrelationships among conditions, cues, and elements of performance.

Directly next to the listing of elements of task performance the designer is asked to document the conditions associated with each step or decision. This documentation is especially important later in test design. Many of you may have already recognized that the Task Structure Analysis represents the highest fidelity test which could be used to measure competency. If an examinee were presented with all possible cues and conditions, performed all steps and decisions, and produced a product with all of the desired characteristics, we could certainly certify his competency. By carefully documenting the conditions associated with this "ideal" performance test we have a basis for development of a simulation which tests performance at the highest practical fidelity.

At the end of the Task Structure Analysis we document the characteristics of the desired results of task performance. When the task does result in a tangible product we are most interested in a listing of the characteristics of that product. Again, this is motivated by our needs in test development. Characteristics of the product are especially important to us because our design methodology stresses the use of product measures whenever possible. When no tangible product is available, a listing of key steps or known errors in the process of task performance may be used. This listing of process steps or errors then becomes the basis for a successful process measure to test competency.

We do not have time today to discuss in detail our methodology for converting the detailed Task Structure Analysis into criterion tests and instructional materials. In general terms, however, a close examination of the cues and conditions of performance will indicate real world constraints which preclude full fidelity testing. These constraints must be overcome either by simulation or, if that course of action is impractical, elements of the total real world performance must be eliminated from the test. In either event the detailed Task Structure Analysis serves as the full fidelity measure from which cost effective criterion tests are developed.

With respect to instructional design, mere examination of each cue, each step and each decision in the process should result in identification of the skills and knowledges inherent in the process and their logical sequence.

In summary we have found it useful to separate Task Analysis activities into two discrete operations.

(1) The preparation of short terse statements written at a level of specificity which not only facilitates data collection for task selection purposes, but also facilitates further detailed analysis and

(2) The detailed analysis of selected tasks into initiating cues, and elements, conditions and results of performance to facilitate criterion test and instructional design.

TASK STRUCTURE ANALYSIS	1. DATE 2 Jul 75	2. PAGE 1 OF 4 PAGES
3. JOB/MOS: MOS: VARIABLE	4. DUTY: First Sergeant	
5. TASK STATEMENT: (STATE AS AN ACTION VERB WITH AN OBJECT.) Select A Detail Using A Duty Roster		6. TASK NUMBER

7. CUES: (LIST EVENTS WHICH INITIATE TASK PERFORMANCE.)	GO TO STEP	NECESSARY CONDITIONS
1. Oral or written requirement to select a recurring detail.	1	# , Date & Type of Detail.
2. Recurring requirement for a detail (SOP).	1	
3. Change in status of anyone on duty roster after publication of detail roster.	24	Notification of Change in Status.

8. DECISIONS AND/OR STEPS: (STATE DECISIONS AS YES/NO QUESTIONS.) (STATE STEPS AS SUBTASKS.)	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
	YES	NO		
1. Do you have a duty roster for this detail?	2	20	X	All Active Duty Rosters.
2. Secure Duty Roster.	X	X	2A	Appropriate Duty Roster, in Files in Office.
2A. Are all columns of the roster already used.	20	2B	X	
2B. Annotate date of detail to next open column.	X	X	3	
3. Are there any personnel to be added to the Roster?	18	4	X	Notes Indicating Required Additions.
a. New arrivals (ASGD or ATCHD).				Notification of Release From ED.
b. Permanent release from ED.				
4. Are there any personnel to be deleted from the roster?	19	5	X	Notes Indicating Required Deletions.
a. Departures (reassignment or Rel from ATCHD).				Notification of New ED.
b. New permanent ED.				
5. Any authorized non-availables? (LV, PASS, SD, TDY, SICK-LINE of Duty).	6	7	X	Notes Indicating Status of Individuals.
6. Post "A" opposite name under date of detail.	X	X	7	
7. Any unauthorized non-availables? (AWOL, SICK-NLD, Confinement, Arrest, other reason due to own misconduct).	7A	8	X	Notes Indicating Status of Individuals.

ADMINCEN FORM 1997, 1 JUL 75
(ATSG-TE)

PREVIOUS EDITIONS MAY BE USED.

TASK STRUCTURE ANALYSIS PAGE 2 OF 4 PAGES	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
	YES	NO		
7A. Post "U" opposite name under date of detail.	X	X	8	
8. Any eligibles who cannot be selected due to previous detail or other duty?	8A	9	X	Other Detail Rosters, Knowledg of Other Duty Requirements.
8A. Post "D" opposite name under date of detail.	X	X	9	
9. Is this a consolidated roster?	10	11	X	Consolidated or Non-consolidated Roster.
10. Select previous column pertaining to category of detail (weekend/holiday or weekday).	X	X	11	
11. Identify (next) highest number in the selected previous column without an "A", "U" or "D" under date of detail.	X	X	12	
12. Is there more than one Soldier with the same highest number?	13	14	X	
13. Does the (remaining) detail requirement equal or exceed those identified?	14	13A	X	
13A. Select sufficient individuals to fill detail requirement by going down from top of roster.	X	X	14	
14. Place hatched lines, in pencil, opposite selected name(s) under date of detail.	X	X	14A	Pencil.
15. Are more individuals required to fill detail requirement?	11	16	X	
16. With the exception of those posted with "A", add 1 to previous column running total and post under date of detail (use red pencil for weekend/holiday columns on consolidated rosters).	X	X	17	Red Pencil. Black Pencil.
17. File Duty Roster and publish Detail Roster (separate tasks).	X	X	EOT	
18. Annotate name to bottom of roster and line out previous detail columns, annotate reason on reverse side.	X	X	4	
19. Delete name from roster and annotate reason on reverse side.	X	X	5	
20. Secure blank Duty Roster.	X	X	21	Blank Duty Roster

TASK STRUCTURE ANALYSIS	PAGE 3 OF 4 PAGES	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
		YES	NO		
20A. Fill in nature of duty, organization, and from date.		X	X	21	
21. Identify all eligibles for entry on roster.		X	X	22	Unit Roster or Previous Duty Roster (Filled)
22. Enter on roster alphabetically by rank.		X	X	2B	
23. Post changes to duty roster.		X	X	24	
24. Is there any change in status of selected individuals in the detail roster which could preclude their pulling the detail?		25	EOT	X	Notification of Change in Status of Individuals in Published Detail Roster.
25. Erase hatched lines pertaining to those individuals and post new status.		X	X	11	Eraser.

PRODUCT CHARACTERISTICS:

1. List of selected individuals for detail
   - a. Proper number.
   - b. Correct names.
2. Properly posted duty roster
   - a. Correct date of detail in column heading.
   - b. "A" posted by appropriate name.
   - c. "D" and correct number posted by appropriate name.
   - d. "U" and correct number posted by appropriate name.
   - e. Hatched lines by appropriate names.
   - f. Correct numbers posted by all other names.
   - g. Correct names added to roster.
   - h. Correct names deleted from roster.
   - i. Proper annotations made on reverse of roster.

TASK STRUCTURE ANALYSIS	PAGE 4 OF 4 PAGES	DECISIONS		GO TO STEP	NECESSARY CONDITIONS
		YES	NO		
j. Correct heading on new rosters. k. Personnel listed alphabetically by rank on new rosters. l. Red entries for weekend/holiday details on consolidated rosters.					

DETERMINING TRAINING PRIORITIES FOR JOB TASKS

By

Lt Col Donald F. Mead, USAF
Occupational and Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland AFB, Texas

I. INTRODUCTION

Few applications of Air Force occupational survey data offer the potential cost savings of establishing or modifying technical vocational training programs. Survey data have broad potential input to the instructional system development (ISD) programs being implemented by each of the military services. A few examples are:

a. Identification of tasks to be included in or eliminated from course curricula.

b. Development of specialty and course training standards.

c. Determining appropriate training emphasis for job tasks.

d. Development of behavioral training objectives.

e. Criterion testing.

Historically the development of a technical training course curriculum has relied on subjective judgments, personal knowledge and experience of subject matter specialists. Recent research by AFHRL has been directed toward the development of a methodology for converting survey data into objective training priority indices for individual job tasks. Research to date has concentrated on the identification of task factors, used as predictor variables to be incorporated into a multiple regression equation yielding a training priority index. Simply stated, the rationale of this research suggests that in making training decisions, curriculum specialists may use specific task information in consistent ways which may be duplicated by mathematical equation. The ultimate objective of this research is to develop a single universal training priority equation which is applicable to tasks in all career ladders, or separate equations for each of the four broad aptitude areas. Success in this endeavor would permit the collection of factor data in conjunction with occupational surveys and the immediate computation of task training priority indices for use by the appropriate training school.

Several task factors used in training decisions were identified in a pilot study by Mial and Christal (1974) using the Medical Services (902X0) specialty. In this study potential factors were isolated and defined and, where appropriate, seven-point rating scales developed. One hundred ninety first-term airman tasks were selected for study and factor ratings for each task obtained from experienced 902X0 NCOs. The task statements and complete task factor information were placed on individual 4" X 6" cards. Complete card decks were administered to curriculum specialists who rank ordered the tasks according to priority for resident technical training. Mean rank values served as the criterion measure and multiple regression analysis was used to "capture" the decision policy of the curriculum specialists. The feasibility of this approach was established in the development of a four-variable equation yielding task values which correlated $R = .88$ with those of the curriculum specialists.

The following four predictor variables made up the training priority policy equation:

a. Task learning difficulty: The time required to learn to perform the task satisfactorily (Mead; 1970, 1970a, 1970b). This variable data is routinely collected in AF occupational surveys.

b. Percent members performing: The percent of airmen in the career ladder performing the task in their jobs. This variable data is routinely computed during the analysis of AF occupational surveys.

c. Consequences of inadequate performance: The perceived consequences if the task is not performed correctly, in terms of destroyed material, wasted time, or loss of life. Factor data were collected via a 7-point rating scale with responses varying from negligible (1) to disastrous (7).

d. Probability of emergency performance: The probability that performance of the task will occur under emergency conditions (those conditions requiring instant and effective action with no time for preparation or practice).

Based on the favorable results of this pilot effort, a follow-on cross validation study was conducted using the Law Enforcement (812XX) specialty. This paper addresses the procedures and findings of this study.

II. PROCEDURES FOR VALIDATION STUDY

Job tasks. The task listing from the Law Enforcement occupational survey was screened to identify apprentice and journeyman level tasks. A representative sample of 165 tasks were selected as candidates for formal vocational technical training.

Predictor variables. The four predictor variables identified in the pilot study were applied to cross-validate their predictive utility. Data for two of the variables, percent members performing and task difficulty, were available from the completed survey data. A listing of the 165 tasks and rating scales for the two remaining variables, consequences of inadequate performance and probability of emergency performance, were administered to thirty 812XX supervisory personnel to obtain data for these variables. Arithmetic means for these variables were computed for each task and used in the analysis as task variable values.

Criterion measures. Thirty-eight experienced raters including instructors, curriculum specialists, ISD specialists, and NCOs provided criterion measures. Individual task cards were prepared listing the task statement and all variable information. Raters were given the 165 task cards and asked to rank the tasks as to priority for formal training. One subgroup of 12 instructors were asked to rank the 165 tasks using only the task statements without the variable information. All raters were requested to use their personal knowledge and professional judgment in ranking the tasks and not to consult specialty or course training standards. After each rater was finished, he was asked to further subdivide his ranked tasks into seven categories using the following definitions:

1. Tasks for which formal training has no value.
2. Tasks for which formal training has little value.
3. Tasks for which formal training has some value.
4. Tasks for which formal training is valuable.
5. Tasks for which formal training is very valuable.
6. Tasks for which formal training is extremely valuable.
7. Tasks for which formal training is absolutely necessary.

Mean rank and category values were computed for each task and used as criterion values in the analysis of the data. In the pilot study the criterion values reflected task priority for resident technical training. Post-administration critiques revealed that for many raters this criterion required two separate decisions: (1) which tasks require formal training, and (2) which tasks require resident training as opposed to OJT. The criterion for the cross validation study eliminated this problem by using task priority for formal training, including both resident and OJT.

Development of training priority policy equations. To capture the training priority policy reflected in the task rankings and ratings, the predictor variables and various derivations from them were analyzed through a series of multiple regression problems. The computed R and R^2 values revealed those predictors and appropriate weights which have the greatest utility in reproducing the training specialists' training priority evaluations.

III. RESULTS

Task variable ratings. Estimates of reliability were computed for the task variables using the intraclass correlation technique (Lindquist 1953, p. 361). The results are shown in Table 1.

Table 1. Estimated Reliability of Task Variable Ratings

Variable	N	R_{11}**	R_{kk}
Consequences of Inadequate Performance	30	.37	.95
Probability of Emergency Performance	30	.24	.87
Task Difficulty	94	.21	.95
Task Delay Tolerance*	120	.44	.99

*See paragraph below.

**Mean reliability for one rater (sample value).

As shown, the variables used, definitions provided, and scales employed resulted in considerable agreement among the raters in evaluating the tasks. In post-administration critiques raters stated that many of the task statements contained the word "emergency" which tended to influence their Probability of Emergency Performance ratings. As a result this variable, for use in later studies, has been changed to Task Delay Tolerance and the scale changed accordingly. In a subsequent study (Goody and Watson, 1975) showed that this variable has a high degree of reliability (see Table 1). Indications are that the variable values employed in the analysis are stable, reliable measures. From these reliability estimates it is apparent that experienced NCOs can provide consistent task variable information using the definitions and scales employed in this study.

Training priority criterion values. Training priority rankings for the 165 tasks were provided by 38 raters experienced in training development. Table 2 provides the inter-rater reliability estimates for the ranking judgments and for the training requirement ratings.

Table 2. Estimated Reliability Coefficients for Training Priority Rankings and Ratings (Criterion Values)

Group	N	R_{11}	R_{kk}
Rankings:			
Total Sample	38	.51	.98
Instructors with Variable Information	12	.61	.95
Instructors without Variable Information	12	.40	.89
Ratings:			
Total Sample	38	.56	.98
Instructors with Variable Information	12	.69	.96
Instructors without Variable Information	12	.44	.90

As noted, there is significant agreement among raters in their training priority ranking and rating judgments. It appears that these judgment decisions are enhanced when task variable information is readily available. The R_{kk} values indicate that the criterion values used in the analysis were stable, reliable measures. Of more significance to the study, these estimates indicate that a consistent training priority judgment policy exists and is available for "capture" via the multiple regression approach (Christal, 1968).

Training priority equation. Each predictor variable, and its squared term where appropriate, was analyzed to develop the multiple regression equation yielding predicted values most similar to the training specialists' training priority evaluations. The training priority rating values served as the criterion. Predictor variables were added in sequential regression problems to determine significant gains or losses in predictive efficiency. The "best fit" policy equation combined the four predictor variables and three derivations shown in Table 3. Variable 6 accommodates those tasks which have high task difficulty ratings (require extended learning time) but are performed by few airmen in the specialty. This equation yielded training requirements indices which correlated $R = .95$ ($R^2 = .91$) with the criterion values.

Table 3. Beta Weights for Selected Predictor Variables

Variable	Beta Weight
1. Percent Members Performing	.99105
2. Probability of Emergency Performance	.38346
3. Task Difficulty	.04577
4. Percent Members Performing, Squared	-.41587
5. Probability of Emergency Performance, Squared	-.11300
6. Task Difficulty if $\leq$ 5% Performing, 0 Otherwise	-.21659
7. Consequences of Inadequate Performance	.03725

$R = .953$ $R^2 = .908$

IV. DISCUSSION

"Task criticality" is frequently cited in training literature as a criterion for the train/don't train decision. Unfortunately, there is little agreement on the definition of this term. Agreement on task training requirements among specialists can be duplicated mathematically using the four variables identified in this study. This approach would provide an objective basis for the curriculum development process. Curriculum specialists could override the computed training priority index but should be required to provide their rationale for doing so.

In the present study, agreement among specialists on task training requirements improved when variable information was provided. On the basis of these findings Air Training Command is initiating the routine collection of variable data in conjunction with occupational surveys. Table 4 shows the format of a newly developed CODAP program which will provide the variable information to aid ISD and curriculum development personnel in their decision-making.

Table 4. Training Variable Information Format

Task Statements Listing	Percent Members Performing	Task Delay Tolerance	Task Difficulty	Consequences of Inadequate Performance

The task listing may be ordered on any one of the variable values. With the development of an operational training priority equation, a fifth column will be added for the training priority index, and tasks will be ordered according to the magnitude of their index value.

The success of the present study has prompted two additional research efforts. In the first, a research contract has been let with Applied Psychological Services to further validate the approach using one career specialty within the specific aptitude areas: electronics, mechanical, general, and administrative. This study will test the utility of the identified variables across career ladders to determine whether additional predictors are required.

The second effort addresses the ultimate objective of this research, to develop a single universal training priority equation. Such an equation would permit the immediate computation of task training priority ratings for each completed occupational survey and require only the gathering of predictor variable information. An essential step in this program is the development of "benchmark" scales to obtain the predictor variable information using common frames of reference for various specialties. The approach and progress of this research is the topic of the paper by Goody and Watson (1975).

REFERENCES

Christal, R.E. JAN: A technique for analyzing group judgment. *The Journal of Experimental Education*, 1968, 36(4), 24-27.

Goody, K. & Watson, W.J. Task Factor Benchmark Scales for Use in Determining Training Priorities. Paper presented at the *17th Annual Conference of the Military Testing Association*, Indianapolis, September 1975.

Lindquist, E.F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin Co., 1953.

Mead, D.F. *Development of an equation for evaluating job difficulty.* AFHRL-TR-70-42, AD-720 253. Lackland AFB, Tex.: Personnel Division, Air Force Human Resources Laboratory (AFSC), November 1970. (b)

Mead, D.F. *Continuation study on development of a method for evaluating job difficulty.* AFHRL-TR-70-43, AD-720 254. Lackland AFB, Tex.: Personnel Division, Air Force Human Resources Laboratory (AFSC), November 1970. (a)

Mead, D.F. & Christal, R.E. *Development of a constant standard weight equation for evaluating job difficulty.* AFHRL-TR-70-44, AD-720 255. Lackland AFB, Tex.: Personnel Division, Air Force Human Resources Laboratory (AFSC), November 1970.

Mial, R.P. & Christal, R.E. The determination of training priority for vocational tasks. *Proceedings, Psychology in the Air Force Symposium.* USAF Academy, April 1974, 29-33.

TASK FACTOR BENCHMARK SCALES
FOR USE IN DETERMINING TRAINING PRIORITIES

By

Squadron Leader K. Goody & Squadron Leader W. J. Watson (RAAF/USAF)
Occupational & Manpower Research Division
Air Force Human Resources Laboratory
Lackland AFB TX 78236

This is the final paper in a series of three dealing with the achievements of the Air Force Human Resources Laboratory (AFHRL) in developing techniques for analzying task factor data and in using such task factor data to determine the training priorities of job tasks. In the second of the series, Mead (1975) points out how AFHRL has identified the four task factors which appear to be the best determinants of priority for formal training. He also indicates how the CODAP programs described in the first of these three papers (Christal & Weissmuller, 1975) can be used to arrange task factor data in a format most suitable for use by training managers in making better training decisions. Finally, Mead establishes the feasibility of using regression equations to compute a training priority index using these task factors and various derivations from them as predictors. The new CODAP programs can then list all the tasks in a specialty ordered from high to low on this index, presenting pertinent task factor data beside each task. The tasks near the top of this list are those on which training is essential, and those near the bottom are the ones with low priority for training. In between these two extremes are all the "could be trained" tasks listed in order of their priority for universal training.

This methodology would have various applications. It could be used for the initial design of a training program, the validation of an existing program, or for the redesign of an existing one given new constraints such as change in course length. At all times it must be remembered that this methodology is an advanced aid to course design, always subject to human override where such can be justified. Perhaps the biggest danger to this project is the potential for blind application of the mathematical model without due consideration for the exceptional case. However, provided training management is aware of this, the model offers vast potential as an aid to optimum training decisions.

The major limitation in the methodology at present rests in the method used for measuring the task factors, apart from the percent-performing factor which is an absolute value. The factors are measured by rating tasks relative to other tasks in a specialty. Specialists consider each task in turn, and then allocate it a rating relative to all other tasks in his specialty. The arithmetic mean of all ratings for a task is used as the measure of that factor for that task. The problem is that the numerical value so ascribed to a task depends on the nature of the other tasks in the specialty. To illustrate, consider two tasks with equal degrees of "Consequences of Inadequate Performance," one a Recruiter task and the other an Operating Room Specialist task. Operating Room Specialist tasks, on the whole, tend to be more consequential than Recruiter tasks. In our example, therefore, although the consequences of failing to adequately perform the two tasks are equal, the Recruiter task will be ascribed a higher numerical value because it is being compared to relatively less consequential tasks. Such ratings can be used to determine an optimum regression equation within a specialty; this has been established (Mead, 1975). However, a new regression equation must be computed for each specialty. The possibility exists that, provided the task factors can be measured on common scales, a limited number of regression equations can be computed, each applying across a number of specialties. In order to examine this possibility, the development of common scales for measuring the factors is being researched, the progress of this research being the subject of this paper.

The goal is to develop a series of "benchmark scales" which should permit measurement of task factors against common frames of reference for various specialties. To illustrate the concept of a benchmark scale, Appendix A contains the tentative benchmark scales for the three factors for those specialties with an administrative or general aptitude requirement. Each comprises 27 tasks, subdivided into nine subgroups of three tasks, each subgroup representing a level on the relevant factor. In operational use, specialist raters will rate each of the tasks in their specialties against the 27 tasks of the benchmark scale rather than against the other tasks of their own specialty. For each task the rater will decide which group of three tasks in the scale has about the same degree of the factor under consideration, the level associated with that group of three being his rating for that task.

In all, nine scales are being developed, one for each of three factors for each of three groups of specialties. The three factors are Consequences of Inadequate Performance, Task Delay Tolerance, and Task Learning Difficulty, definitions of each being included in Appendix A. The three groups of specialties are defined on the basis of aptitude requirements: the first contains specialties with an administrative or general requirement, and the other two groups are those with an electronic or mechanical aptitude requirement respectively.

The development of a benchmark scale is a three-phase process. Initially, data is gathered and analyzed for a reasonably large inventory of tasks. This data is then used to select the 27 tasks for the tentative benchmark scale; and finally the tentative scales must be validated. After elaboration on this process, results obtained to date will be discussed.

The initial inventory contains tasks typically performed by a variety of specialties, but each must be of such a nature as to be meaningful to experienced airmen from other specialties. For example, the inventory for specialties with an administrative or general aptitude requirement contained a total of 438 tasks representing 54 different specialties. As the present concern is for initial trade training, only journeyman-level tasks are used. The inventory must represent the full range of values on each of the factors for which it is used, and multi-facet tasks consisting of several divergent components are avoided. The specialty normally associated with each task is indicated in the inventory.

For each factor, about 120 raters rate each inventory task, on a 9-point scale, relative to the other tasks in the inventory. The raters are selected randomly from the first-line supervision level (7-skill level) of the group of specialties under consideration, different raters being used for each factor. This is necessary to avoid spurious interaction between factors; particularly as Task Delay Tolerance, with Level 1 the most demanding situation, is effectively an inverted scale relative to the other two. Such large numbers of ratings are necessary to provide confidence in the stability of the means obtained by having raters rate tasks from other specialties. First-line supervisors were selected as raters in order to obtain the optimum blend of general experience and first-hand knowledge of journeyman-level activities.

The raw ratings are then refined and arranged into a format suitable for selecting the tasks for each benchmark scale. A standard computer program (Stacey, Weissmuller, Barton, & Rogers, 1974) identifies the occasional rater who has not taken his task seriously, and he is eliminated from further processing. Some raters rate harshly, others are more lenient. To compensate, all the ratings for each rater are standardized to a common mean of 5.0 and standard deviation of 1.0. The mean and standard deviation of these adjusted ratings for a task are then taken respectively as the most appropriate measures of the relevant factor for that task and of rater agreement on that measure.

For each factor, the tasks are then ordered on the mean rating, this listing being used to select the nine groups of three tasks for the benchmark scale. All three tasks representing one level must lie near each other on this listing, but be relatively distant from the groups chosen to represent the levels on either side. The magnitude of the difference in mean rating between successive groups must be uniform, and preference is given to tasks with the lowest standard deviations--this indicating the best rater agreement. Tasks selected must provide maximum coverage of the various specialties, and they must be those most likely to be meaningful to airmen from other specialties. Finally, similarities within groups and differences between groups must be apparent as well as measured.

Finally, in the validation phase, the tentative benchmark scales are tested to determine whether raters can indeed rate tasks against them, and whether such ratings are independent of the general level of the factor within the specialty. This is done by selecting tasks from the original inventory, having them rated by small samples of raters, and comparing the results with the ratings gathered for the full inventory. The tasks selected from the original inventory are all those representing about 10 selected specialties. A general sample of raters rate all these tasks against the benchmark scale, and specialist raters rate those from their specialties only (4-20 tasks). For each of these tasks, there will thus be three sets of ratings: the original ratings, ratings against the benchmark scales made by general raters presented with the full subset of tasks, and ratings against the benchmark scales made by specialists presented only with the few tasks from their own specialty. Analysis of the means, standard deviations and inter-correlations of these sets of ratings will indicate the usefulness of the scales.

The three scales for specialties with administrative or general aptitude requirements are now in the validation phase. One aspect of concern was whether supervisors could rate tasks from other specialties. As can be seen from the inter-rater agreement coefficients in Table 1, this is no longer a concern; such raters can rate tasks from other specialties with a high degree of agreement.

Table 1. Inter-rater Agreement Coefficients

	N	R_{11}	R_{kk}
Consequences of Inadequate Performance	116	.51	.99
Task Delay Tolerance	120	.44	.99
Task Learning Difficulty	117	.48	.99

A series of correlation coefficients were computed to examine the relationships between rating means and standard deviations. These correlation coefficients are presented in Tables 2, 3, and 4. Correlations were computed between the adjusted task means and standard deviations for each factor (Table 2). The existence of a significant correlation would indicate a relationship between rater agreement and mean rating. There is evidence that the mean and standard deviation for Task Delay Tolerance are positively related, suggesting that the raters showed more agreement on those tasks with low task delay tolerance. Table 3 is the correlations between the means for the three factors, and Table 4 between standard deviations. Significant between-factor correlations exist in all cases. The same tasks tend to be

Table 2. Correlation Between Task Means and Standard Deviations

Consequences of Inadequate Performance	.01
Task Delay Tolerance	.22
Task Learning Difficulty	.13

Table 3. Correlation Between Task Means

	A	B	C
A. Consequences of Inadequate Performance	1.00		
B. Task Delay Tolerance	-.84	1.00	
C. Task Learning Difficulty	.58	-.38	1.00

Table 4. Correlation Between Task Standard Deviations

	A	B	C
A. Consequences of Inadequate Performance	1.00		
B. Task Delay Tolerance	.53	1.00	
C. Task Learning Difficulty	.58	.27	1.00

rated towards the more demanding end of each scale, and rater agreement depends to some extent on the task and/or its statement regardless of the factor being rated. The high correlation between the mean ratings on the first two factors is understandable as delayed performance is one form of inadequate performance, and the delay that can be tolerated is to some extent a function of the consequences of such delay. However, despite the strong relationship between these two factors, it is believed each has a sufficiently high unique component to contribute significantly to the training priority index model.

For the validation phase of the administrative/general series, a total of 127 tasks representing 13 specialties were selected from the original inventory. Specialists have rated the tasks from their own specialty against the benchmark scales, and general raters have rated the full list of 127 tasks against the scales. As all returns are not in at the time of this writing, the results of the validation phase cannot be included in this report.

The development of benchmark scales for the electronic and mechanical series is not this far advanced. This has allowed the experience gained in collecting the data for the first series to be used to refine the instructions and scales used in the survey booklets. The survey material for the initial data-gathering phases has been prepared and will be administered shortly.

To date, therefore, the Air Force Human Resources Laboratory has been able to identify and define the task factors which are strong determinants of task training priority. Measures of these factors can be presented in a format facilitating training decisions; and, within a specialty, these measures can be used as predictors in a regression equation to compute a training priority index. Research is well advanced into a technology for measuring these task factors against common frames of reference which apply across career fields. On successful completion of this phase, further research will seek a limited number of general regression equations for computing training priority indices. Should this prove successful, training managers will have an economical, efficient, and reliable tool for making optimum training program content decisions.

REFERENCES

Christal, R. E. & Weissmuller, J. J. New CODAP Programs for Analyzing Task Factor Information. Paper presented to 17th Annual Conference of the Military Testing Association, U.S. Army, Indianapolis, IN, 16-19 September 1975.

Mead, D. F. Determining Training Priorities for Job Tasks. Paper presented to 17th Annual Conference of the Military Testing Association, U.S. Army, Indianapolis, IN, 16-19 September 1975.

Stacey, W. D., Weissmuller, J. J., Barton, B. B., & Rogers, C. R. *CODAP: Control Card Specifications for the UNIVAC 1108*. AFHRL-TR-74-84, AD-A004 085. Lackland AFB, Tex.: Computational Sciences Division, Air Force Human Resources Laboratory, October 1974.

APPENDIX A

CONSEQUENCES OF INADEQUATE PERFORMANCE

DEFINITION

Consequences of inadequate performance is a measure of the seriousness of the **probable** consequences of inadequate performance of a task. It is measured in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours of work, etc.

BENCHMARK SCALE

Level 1 – Least Serious Consequences of Inadequate Performance

Deliver newspaper to local distribution points (Information Specialist AFSC 79150)
Clean flight planning room (Command and Control Specialist AFSC 27450)
Fold or count hospital linen (Medical Service Specialist AFSC 90250)

Level 2

Tatto identification on Air Force working dogs (Veterinary Specialist AFSC 90850)
Arrange, mark and display property for best sales results (Materiel Facilities Specialist AFSC 64750)
Enter daily work assignments on time cards (Administration Specialist AFSC 70250)

Level 3

Draw up work rosters for taxi operators or drivers on large AF base (Programs and Work Control Specialist AFSC 55530)
Compute selling price for processed meat and meat produce (Meatcutter AFSC 61250)
Compute quantity of earth to be removed or used for fill (Geodetic Surveyor AFSC 22250)

Level 4

Finish and polish gold alloy inlays, crowns or fixed partial dentures (Dental Laboratory Specialist AFSC 98250)
Operate keypunch machine to keypunch data cards (Personnel Specialist AFSC 73250)
Perform normal satellite photography sequence (Aerospace Control & Warning Systems Operator AFSC 27650)

Level 5

Reload computer after power failures or fluctuations (Communications Center Specialist AFSC 29150)
Detect theft of money or stock from commissaries or supply service outlets (Supply Services Specialist AFSC 61150)
Measure and record auditory acuity or hearing sensitivity (Aeromedical Specialist AFSC 90150)

Level 6

Quell disturbances involving military personnel (Security Specialist AFSC 81150)
Prepare aircrew navigation kits (Air Operations Specialist AFSC 27150)
Take and record pulses, temperatures and respirations (Medical Service Specialist AFSC 90250)

Level 7

Fit cargo parachutes to airdrop cargo (Aircrew Life Support Specialist AFSC 92250)
Analyse radarscope photographs to identify targets or evaluate target condition (Imagery Interpreter AFSC 20650)
Sterilize surgical instruments or supplies (Operating Room Specialist AFSC 90252)

Level 8

Apply first aid at scene of accident or incident (Security Specialist AFSC 81150)
Render missile safe for maintenance or verify missile safing (Missile Safety Specialist AFSC 24150B)
Alert direction finding (DF) stations when aircraft emergencies occur (Radio Operator AFSC 29353)

Level 9 – Most Serious Consequences of Inadequate Performance

Defend AF installations against attack by hostile forces or sabateurs (Security Specialist (Military Dog Qualified) AFSC 81150A)
Assist patient to maintain proper airway during surgery (Operating Room Specialist AFSC 90252)
Rescue personnel from aircraft or aerospace vehicle (Fire Protection Specialist AFSC 57150)

USE OF THE SCALE

1. For each task in turn, think of the **probable** consequences of inadequate performance. Think in terms of possible injury or death, wasted supplies, damaged equipment, wasted man-hours or work, etc.
2. Decide which set of three tasks in the above scale have about the same consequences of inadequate performance.
3. The level indicated for this set of three tasks is your measure of the consequences of inadequate performance for the task under consideration.

APPENDIX A

TASK DELAY TOLERANCE

DEFINITION AND EXPLANATION

The Task Delay Tolerance of a task is a measure of how much delay can be tolerated between the time the airman becomes aware the task is to be performed and the time he must commence doing it. Must he commence immediately, or does he have time to consult a manual, seek guidance, or even be taught how to do it?

BENCHMARK SCALE

Level 1 – Least Tolerance of Delay

Use artificial respiration to restore breathing of accident or fire victims (Fire Protection Specialist AFSC 57150)
Issue scramble orders to fighter aircraft (Command and Control Specialist AFSC 27450)
Assist during treatment of cardio-respiratory failure in operating room (Operating Room Specialist AFSC 90252)

Level 2

Quell disturbances involving military personnel (Security Specialist AFSC 81150)
Identify tablets, capsules or liquids involved in poisoning cases (Pharmacy Specialist AFSC 90550)
Operate safety console at missile control center during hazardous operations (Missile Safety Specialist AFSC 24150B)

Level 3

Inspect runway for foreign objects (Air Operations Specialist AFSC 27150)
Administer anaesthesia in dental surgery (Dental Specialist AFSC 98150)
Adjust airborne radio receivers to obtain readable signals (Radio Operator AFSC 29353)

Level 4

Question suspects or witnesses (Security Specialist AFSC 81150)
Perform colony counts on bacteria to estimate type and level of infection (Medical Laboratory Specialist AFSC 90450)
Maintain proper temperature of food storage areas (Cook AFSC 62250)

Level 5

Identify military vehicles, installations or activities in visual photographs (Intelligence Operations Specialist AFSC 20450)
Proofread or correct teletype tape or page copies (Communications Center Specialist AFSC 29150)
Prepare daily weather maps (Weather Forecaster Specialist AFSC 25330)

Level 6

Operate computer remote inquiry terminals (Computer Operator AFSC 51150)
Purge or clear chemical lines in film developing machines (Still Photographic Laboratory Specialist AFSC 23354)
Service and maintain dental high-speed drilling equipment (Dental Laboratory Specialist AFSC 98250)

Level 7

Monitor workload reporting systems (Manpower Specialist AFSC 73330)
Brief personnel on state or local motor traffic laws (Safety Specialist AFSC 24150)
Draw up work rosters for taxi operators or drivers on large Air Force base (Programs and Work Control Specialist AFSC 55530)

Level 8

Write item identification descriptions and specifications for catalogues (Procurement Specialist AFSC 65150)
Interview or hire civilian personnel (Supply Services Specialist AFSC 61150)
Prepare and analyse work flow process charts (Management Engineering Specialist AFSC 73331)

Level 9 – Most Tolerant of Delay

Review or select books or publications for unit library (Administration Specialist AFSC 70250)
Research and write feature stories in Air Force publications (Information Specialist AFSC 79150)
Clean teeth of animals (Veterinary Specialist AFSC 90850)

USE OF THE SCALE

1. For each task in turn, think of the amount of time that could normally be allowed to elapse from the time the airman who has to do the task first becomes aware of the need to perform it and the time he must actually commence doing it. Think in terms of time he could spend reading up on the task or seeking advice on how to do it, or just thinking about how to do it.
2. Decide which set of three tasks in the above scale have about the same amount of task delay tolerance.
3. The level indicated for this set of three tasks is your measure of the task delay tolerance of the task under consideration.

APPENDIX A

TASK LEARNING DIFFICULTY

DEFINITION

Task Learning Difficulty is a measure of the need for lengthy, systematic training before a new member of the appropriate Air Force specialty could perform the task adequately. It may be thought of as the difficulty involved in "picking up" the task on the job without any systematic training.

BENCHMARK SCALE

Level 1 – Least Difficult to Learn

Collect food trays or serving units from patients in hospital wards (Medical Service Specialist AFSC 90250)
Clean display cases, furniture or fixtures in commissary (Supply Services Specialist AFSC 61150)
Stamp time of receipt on incoming messages (Communications Center Specialist AFSC 29150)

Level 2

Distribute administrative orders within unit (Administration Specialist AFSC 70250)
Schedule health examinations for meat cutting personnel (Meatcutter AFSC 61250)
Count property in warehouse bins or shelves (Inventory Management Specialist AFSC 64550)

Level 3

Challange or identify unknown persons in vicinity of correctional facility (Corrections Specialist AFSC 81251)
Deliver passenger manifests and allied documents to international border clearance authorities (Air Passenger Specialist AFSC 60550)
Act as armed escort for personnel transferring funds (Security Specialist AFSC 81550)

Level 4

Complete and submit Radiation Exposure Registration Form (AF Form 1520) (Environmental Health Specialist AFSC 90750)
Assemble shelter manager kits (Disaster Preparedness Specialist AFSC 24250)
Maintain imprest or petty cash account (Procurement Specialist AFSC 65150)

Level 5

Verify labels or instructions for handling radioactive substances (Materiel Facilities Specialist AFSC 64750)
Inspect buildings for termites or other wood destroyers (Entomology Specialist AFSC 56650)
Prepare comparative productivity charts for work centers (Management Engineering Specialist AFSC 73331)

Level 6

Control or extinguish structural fires (Fire Protection Specialist AFSC 57150)
Calculate number or amount of each food item to be prepared for therapeutic diet (Diet Therapy Specialist AFSC 62251)
Prepare injured personnel for evacuation by litter or ambulance (Aeromedical Specialist AFSC 90150)

Level 7

Administer Intermittent Positive Pressure Breathing (IPPB) therapy (Medical Service Specialist AFSC 90250)
Analyse computer stops for possible hardware malfunction (Supply Systems Specialist AFSC 64850)
Determine position of aircraft by analysis of radarscope photographs after mission (Imagery Interpreter Specialist AFSC 20650)

Level 8

Operate safety console at missile control center during hazardous operations (Missile Safety Specialist AFSC 24150B)
Differentiate between actual targets and electronic countermeasures or decoys (Electronic Warfare Countermeasures AFSC 27631)
Determine axis of attack for air-to-ground attack missions (Intelligence Operations Specialist AFSC 20450)

Level 9 – Most Difficult to Learn

Determine chemical composition of foreign made drugs (Pharmacy Specialist AFSC 90550)
Perform deep roentgen therapy on tumor or cancer patients (Radiology Specialist AFSC 90350)
Assist during treatment of cardio-respiratory failure in operating room (Operating Room Specialist AFSC 90252)

USE OF THE SCALE

1. For each task in turn, think of its learning difficulty; think in terms of the need for lengthy, systematic training, or alternatively in terms of the difficulty in "picking up" the task on the job without systematic training.
2. Decide which of the three tasks in the above scale have about the same learning difficulty.
3. The level indicated for this set of three tasks is your measure of the learning difficulty of the task under consideration.

The Ultimate Catastrophe

LCDR Clarence L. WALKER, USN
Central Test Site for Personnel and
Training Evaluation Program (PTEP)

The title of the paper is not meant to indicate that we are headed for the "Ultimate Catastrophe". My concern is that, despite contrary indications, we train people to repair equipment at the point of worst hypothesized problem or ultimate catastrophe.

Occupational Analysis addresses task elements which are a commonly employed part of a job and whose frequency and duration can be measured. It also addresses "infrequently performed, high criticality" task elements. These elements are based less on observation than they are on speculation. The equipment design maintenance philosophy and the technician's "If this happened you would have to----" contribute to this speculation. These infrequently performed task elements constitute a large portion of any advanced training provided on an equipment and may well be the determinent of the entry level ability set for the job.

The Fleet Ballistic Missile Personnel and Training Evaluation Program (PTEP) serves as a feedback loop in an instructional system. As such it measures and evaluates personnel and training courses for compliance with and support of job descriptions and evaluates job descriptions for correlation with performance. Evaluating an advanced training course in the summer of 1974 PTEP found a significant difference between the difficulty level of material being presented and the level required by the curriculum in support of our task elements. Believing that, the best efforts of training managers aside, instructors teach what is needed to support the equipment; an investigation was undertaken of school graduate performance in the field. Our intent was to determine if school graduates used the skills they had learned to maintain equipment more successfully than non-school graduates maintained the equipment. To this end, PTEP launched into a reduction of equipment maintenance data for correlation studies with training data. The analysis included types of faults occurring, frequency of faults and fault isolation procedures required to effect repairs. This type of analysis was next employed in evaluation of a second course. The results of just these two evaluations were of sufficient interest that an evaluation of the maintenance training task inventories in support of two entire Navy Enlisted Classifications (NEC) was undertaken. Table (1) below summarizes in general terms these evaluations:

1. MTRE MK 6 MOD 3

Course Analysis - Instruction provided was below the level proposed to support task elements.

Equipment Maintenance Data Analysis - Four years of failure reports showed no failures requiring the advanced level of knowledge proposed by the task inventory.

2. MK 3 MOD 0 Guidance Assembly

Course Analysis - Instruction was provided at the level proposed by the task inventory.

Equipment Maintenance Data Analysis - Information provided in the course was not employed in actual fault isolation procedures.

3. Central Navigation Computer (CNC) Technician (NEC 3333)

Equipment maintenance data analysis showed significant differences between the circuit and computer theory taught in support of the task inventory and that employed in fault isolation and repair.

4. Navigation Data Assimulation Computer (NAVDAC) Technician (NEC 3332).

Equipment maintenance data analysis indicated the task inventory accurately reflected fault isolation requirements.

The underlying problem in each of the inaccurate inventories was "infrequently performed, high criticality" task elements which reflected fault isolation requirements for electronic equipment of the mid 60's vice early 1970's and the tendency to support equipment to the "Ultimate Catastrophe". The NAVDAC is an early 1960's equipment). In each case the accuracy or inaccuracy of the task element analysis was objectively demonstrated by reduction of equipment maintenance data. Task elements listed in the task analysis as "infrequent" but highly critical were actually improbable. To support these "improbable" tasks training courses were longer and more complex than they had any need to be. We had worked from what we thought the job was rather than using available data to see what it really was.

Analysis of equipment maintenance data is a powerful and objective tool for bringing reality to task elements and training objectives and as such should regularly be made a formal part of occupational analysis. It can moderate the influence of the tendency we all have to upgrade the difficulty of our jobs and the amount we must know to do them. This tendency markedly affects the present task inventories garnered from questionaries and interviews. Where systems are redundant and survival itself is not at stake, equipment maintenance data may provide justification for reduction in the complexity of training provided or proposed. This last is important if we are to efficiently use the tightening resources provided for training. We have enough real maintenance problems to prepare for; let's work on them and let the "Ultimate Catastrophe's" be.

Samuel Bowser
NPRDC, SDiego, CA

Measurement of Unit Effectiveness: A Criterion Development Problem

Introduction

The U. S. Navy is interested in new ideas and approaches to help it maintain required military effectiveness in light of many human resource problems. The research community has been pursuing research and development projects addressing operational problems in the human resource area for a number of years. The research in this area has demonstrated the lack of adequate and/or usable criteria for determining unit effectiveness.

The search for criteria to measure (unit) organizational effectiveness is complex. A variety of methods for identifying and studying criteria are evident in the literature. Approaches include the multitrait-multimethod matrix, cost-benefit ratios, systems analysis, and criterion composites. None of them have come close to having complete or thorough answers. It would appear that research aimed at finding the best group of criteria for many situations will be in vogue until a framework or structure for criteria can be developed.

There has been considerable interest in organizational effectiveness and the criteria necessary to measure it. Research in many different type of organizations will be needed to find overall measures of effectiveness that will generalize. The literature on this subject is quite ambiguous and extensive.

The concept of effectiveness is addressed in a review report by Campbell, Bownas, Peterson, and Dunnette published July 1974 (TR 75-1). This report outlines two general models of organizational effectiveness. The first, known as the "goal centered" view, defines effectiveness as progress toward and/or achievement of organizational goals. Effectiveness is measured by how close the organization comes to meeting its goals. The system output variables are criteria from this point of view. A major problem is defining explicit goals. The other major approach to effectiveness is the "natural systems" model. This view holds that, since organizations are dynamic and complex, it is unreasonable and simplistic to try to determine goal progress. Rather, the flexibility and survival of the organization are critical. These represent the process of the system, given this point of view internal consistency is related to organizational effectiveness.

The various approaches to developing effectiveness criteria assume the existance of an underlying systems model of organizations. Such systems models have been outlined by Katz and Kahn (1966), Blake and Mouton (1968), and Kaufman (1972), among others. These applications of systems methodology are not rigorous as in mathematical systems models. Rather, they serve as useful and realistic frameworks for discussing organizational functioning.

The logical approach to evaluating a system would seem to be to analyze each aspect of the system separately. The input of any given system is the output of some other system. The process variables measure the status of the system during any given time interval. The output variables are measures of products and direction of the system. Since all these aspects are essential to the system's continued operation, it follows that they are all related to organizational effectiveness. The environment in which the system functions is also an integral part of the evaluation of the system.

This investigation assumed that the starting point for criteria and development was to obtain information about the current evaluation system. The managers of four operational units representing two levels in organizational hierarchy were interviewed. The interviews were content analyzed and the results of the content analysis are the subject of this report.

Method

Subjects: A total of 68 management level persons (67 males, 1 female) from four Navy units participated in the interviews. The ages of the respondents ranged from 23 to 50 years. The subjects were selected based on their job position within their unit. A representation of the managerial structure was sought by selecting a cross section of each units hierarchy ranging from Commanding Officer to leading Chief.

Interviews: The interviews were conducted by professional research associates of the author (6 males, 1 female). The C. O.'s of each unit were interviewed by the one military interviewer, the remaining subjects were randomly assigned to the interview team.

The Session: The subjects below the C. O. were informed of the scheduled interviews. No additional incentive to cooperate was used. Respondents were assured of the anonymity of their interview session and were told that the tape recorder was used because it was the most accurate method of collecting the data. To promote an unstructured atmosphere, subjects were encouraged to make comments, raise questions or discuss impressions at any time during the interview.

The interviewers exercised considerable freedom in rearranging the question order to flow with the prevailing discussion topic. Questions were omitted if the interviewer felt the respondent had elaborated sufficiently on the subject in conjunction with a previous question. At the end of the 30-60 minute interviews, the respondents were asked to fill out and return a biographic questionnaire at their convenience.

Missing Data: The information on several interview tapes was lost due to either malfunctioning recording equipment or excessive background noise.

Scoring Technique: From each of (67 C-90) the cassette tapes of the interviews, transcripts were typed by an independent transcribing service. A data analyst verified the accuracy of the transcription by reviewing the manuscript and tape concurrently. In addition to a direct listing of responses, at this point it was possible to interpret factitious or connotative intent of the respondents and thus modifying the written copy. Interpretation was made due to the intonation of the respondents voice.

Content Analysis: After the responses from the first unit (N=25) were listed, they were categorized with the following parameters given consideration:

(1) To which component of an input-process-output system model does the criteria coincide?

(2) Is this criteria objective (quantifiable) or a subjective measure?

(3) Does this criteria apply to an individual, an organization, or both?

Within categories, several other logical conventions were employed such as ordering criteria from general to specific and from formal to informal. As additional interview tapes and manuscripts were scored, the categories expanded with greater detail.

Questionnaire: The structured interview questionnaire included thirty-one questions. There were 16 questions which apply to this report. These questions can be broken down into three areas and two types. The areas include questions about the unit as a whole, the departments within the unit and the goal of the unit. Each of these areas has questions concerning ideal (should be) and applied (as is) criteria and methods.

Unit Questions from Interview

Applied

17. What are the most important factors in making this unit effective?

19. What are the principal criteria used to evaluate this unit?

Ideal

11. What are your suggestions for alternative methods of evaluating this unit?

13. If you were asked to implement an evaluation system for this unit, what method would you use?

16. In your opinion, what factors make an organization effective?

21. What criteria in your opinion should be used to evaluate this unit?

28. What factors should be related to the effectiveness of this unit?

Departmental Questions from Interview

Applied

5. What methods are used to evaluate the departments of this unit?

20. What criteria are used to evaluate the departments of this unit?

29. What are the most important factors in making this an effective department?

Ideal

12. What methods should be used in your opinion to evaluate the departments/divisions of this unit?

Goals Questions from Interview

Applied

23. What are the goals of this unit?

24. What methods of evaluation are there that assess goal achievement?

26. What kinds of goals are there that relate your position in the organization to the units goals?

Ideal

22. What goals should this unit have in order to best perform its mission?

25. What evaluation methods could be used that would assess goal achievement?

Results

A summary of the content analysis categories is presented in Table 1. This table also gives the number of individuals who responded to each category over all the interview questions; also the total number of responses is reported. The categories of response were rank ordered by number of respondents, the results for the first twenty-seven items are presented in Table 2. The results indicate that first of all military managers are immediate performance oriented.

The overall results of the interviews give "meeting operational commitments" as the most important evaluation criterion perceived by the respondents. This response category was used at least once by 83.8% of those interviewed. The next closest category was efficiency which was given by 47.1% of the respondents. When analyzed on a unit basis, the operational units used the category "meeting operational commitments", however, the staff level group did not; they accounted for only 5.2% of the total response in this category. The staff level unit in my opinion is more concerned with how the job is done, however, the operating level managers perceive immediate performance as the main criterion. There is a potential for non stated conflicting goals between these levels. The next section of the results will give details of the top several responses. The interpretation of the answers by question area (Unit, Dept., Goals) and type (Ideal, Applied) will be given.

Meeting Operational Commitments: This response was seen by 54% as the method and criterion by which evaluation is accomplished. This was also given as a reason for the unit being effective (40%), but when asked about organizational effectiveness in general, only 21.7% responded this way. When questioned about what kind of evaluation system they would use, only 11.7% stated this category would be used as a method.

The criteria (20%) and methods (16.7%) used to evaluate the departments of this unit is meeting operational commitments. When asked what method should be used, 12% stated this category of response, but only 5.1% stated it was an important factor in making the department effective. There was agreement as to the operational value of this criterion but little endorsement of its theoretical value.

When questioned concerning the goals of the unit, meeting operational commitments, was used by 47.9% of those responding to this question. When questioned about evaluation of goal achievement this criterion was used by only 20% of those responding. It would seem that this area is not being used for goal evaluation in most respondents perception.

Efficiency (monetary-manpower): This criteria was endorsed by 20.8% as an area which should be related to unit effectiveness and by 13.3% as being related to organizational effectiveness in general. It was seen as a criterion which should be used to evaluate the unit by 14.0%. Only 4.4% saw this criterion as one which is an important factor in making their unit effective. Thus, this area is related to theoretical effectiveness but not in "our" case.

This criterion is seen as being used as a criterion in relation to the departments by 7.5% of those responding to the question. While this is not a large response it is larger than those who think this should be used as a criterion (2.6%). There is an element of rejection of this factor as an evaluation criterion in the case of departments.

The area of goals is the strongest factor of endorsement, with 31.3% viewing this criterion as an evaluation method of choice for the unit. The aspect of this being a unit goal was played down, only 12.5% used this content factor.

Training: This criterion is related to organizational effectiveness (23.3%) and should be related to unit effectiveness (14.6%), however, it is not now used as a criterion (2.2%). In fact, it should not be used since only 2.0% indicated it was and no one said they would use it to evaluate the unit.

The departmental level had the same view as the unit, that is, training is related to theoretical effectiveness but not in this individual context. This is also true in goals, training is a unit goal (16.7%), but don't evaluate it. The indication is that the operational managers don't want this area explored for evaluation purposes.

It is also a fact that on several of the criterion variables there were differences by rank, age, education, etc. The number of respondents is small enough that interpretation is not advisable, however, future research will give this due consideration.

Conclusions

This data represents the perceptions of operational managers. The areas of criteria identified will move us toward more complete evaluation procedures than we now engage in.

Table 1

Content Analysis Categories

		Number of Individuals Using this Category	Number of Times this Response was Used
1.	General evaluation criteria (This includes the following: overall view of the evaluator; details; personnel judgment or opinion of the evaluator; first impression of the evaluator; chance or haphazard factors; method of performance; results of performance.)	29	67
2.	Required Self Report (i.e., 3M, NAVFORSTAT)	11	16
3.	Exercises and/or drills (This includes the following: personnel evaluations, fitness reports; fleet exercises.)	19	27
4.	Inspections and Audits (i.e., ADMAT, Safety)	14	23
5.	Positive (feedback, critical incident)	15	25
6.	Negative (feedback, critical incident)	17	2
7.	Neutral (survival, don't rock the boat)	4	5
8.	Biodemographic (This includes the following: potential IQ, health and physical ability; education; experience; problem solving ability.)	20	25
9.	Technical proficiency (knowledge, skills)	17	25
10.	Personality (initiative, appearance, general quality)	30	40
11.	General human relations	9	10
12.	Morale	29	58

13.	Satisfaction	9	9
14.	Pride	10	17
15.	Teamwork (absence of conflict, personality or otherwise)	15	11
16.	Attitude (includes cooperation and independence)	22	38
17.	Motivation (do a good job daily)	14	16
18.	Dedication	11	13
19.	Discipline (This includes the following: discipline in general; Captains Mast; racial problems; following orders; respect for seniors.)	10	18
20.	Leadership (This includes: general leadership; inspiring supervision; effectiveness; subordinate empathy, warmth toward men)	20	31
21.	Management (This includes: professionalism; planning; foresight.)	17	18
22.	Awareness of relation of task to mission (the larger view)	16	22
23.	Reports (general, accurate, on time)	19	33
24.	General Administration (This includes: discipline actions; message traffic; quizzes; history of the unit; public relations.)	17	23
25.	Communications network (dissemination of information)	20	30
26.	Adherence to bureaucratic standards (This includes: general; systematic interpretation of standards; records.)	20	33
27.	Paperwork	14	18
28.	Sufficient resources (parts, adequate equipment procurement)	27	33

29.	Quality of equipment	16	32
30.	Adequate maintenance - services support	17	21
31.	Physical environment (This includes: adequate work space; effective layout; working conditions; orderliness; cleanliness.)	20	28
32.	Personnel and manning (This includes: general manning level; size of unit—small enough or large enough; individual task specialization; individual task versatility; human resources utilization.)	23	33
33.	Personnel availability (This includes: turnover vs. continuity; advancement rates; reenlistment rates.)	10	20
34.	Meeting operational commitments	57	191
35.	Operational readiness (assigned tasks)	28	46
36.	Operational capability (combined with mission availability)	14	19
37.	Operational time factors (adherence to schedules; turn around time)	24	28
38.	Training (This includes: quality training programs; training aids; training time.)	31	64
39.	General efficiency (monetary - manpower)	32	60
40.	Safety	21	38
41.	Surprise inspections	14	20
42.	Scheduled inspections (more often - less often)	9	11
43.	Timing of inspections (This includes: amount of warning given in hours)	23	24
44.	Evaluate personnel (management, workgroup, supervisory)	6	6
45.	Evaluate performance (readiness; end product; how job is done)	17	23

46.	Evaluate physical assets (condition of equipment)	3	3
47.	Self inspection for evaluator	9	13
48.	General other evaluator	9	13
49.	Next up chain of command as evaluator	11	11
50.	Similar unit (sister unit) as evaluator	11	12
51.	Customers as evaluators	2	2
52.	X.O. as evaluator	4	6
53.	C.O. as evaluator	[illegible]	5
54.	Experts as evaluators	10	13
55.	Competition for evaluation	7	8
56.	Check-off list	19	22
57.	Observation (in-person evaluation)	7	9
58.	Past performance - history of unit	10	11
59.	Operational assessment (on-going)	18	24
60.	Informal methods of evaluation	7	9

Table 2

Rank Order of First 27 Categories of Response
by Number of Individuals Responding

Rank	Category	N Using Response	% of N Using Response	No. of times Response Used
1.	Meeting operational commitments	57	83.8%	191
2.	Efficiency (monetary - manpower)	32	47.1	60
3.	Training	31	45.6	64
4.	Personality (initiative/appearance)	30	44.1	40
5.	General human relations	29	42.6	58
6.	General evaluation criteria (subjective)	29	42.6	47
7.	Operational readiness	28	41.2	46
8.	Sufficient resources	27	39.7	33
9.	Operational time factors	24	35.3	28
10.	Personnel and manning	23	33.8	33
11.	Timing of inspections	23	33.8	24
12.	Attitude	22	32.4	38
13.	Safety	21	30.9	38
14.	Adherence to bureaucratic standards	20	29.4	33
15.	Leadership	20	29.4	21
16.	Communications	20	29.4	30
17.	Physical environment	20	29.4	28
18.	Biodemographic factors	20	29.4	25
19.	Reports	19	27.9	33
20.	Exercises - drills	19	27.9	27
21.	Check-off list	19	27.9	22
22.	Negative (feedback - critical incident)	17	25.0	28
23.	Technical proficiency	17	25.0	25
24.	Evaluate performance	17	25.0	23
25.	Adequate maintenance	17	25.0	21
26.	Management	17	25.0	18
27.	Quality of equipment	16	23.5	32

Response	Unit Related Questions						
	Applied		Ideal				
	#17 Unit Effective	#19 Eval Criteria	#11 Altern Method Eval	#13 Method You Would Use	#16 Organ Effective	#21 Criteria Should Use	#28 Factors Should Be Unit Eff
Meeting Operational Commitments	40.0	54.9	-	11.3	21.7	54.0	22.9
Efficiency (monetary-manpower)	4.4	7.8	-	9.4	13.3	14.0	20.8
Training	2.2	2.0	-	0	23.3	6.0	14.6
Personality (initiative-appearance)	22.2	2.0	-	1.9	18.3	0	10.4
General Human Relations	4.4	0	-	0	10.0	0	0
General Evaluation Criteria	13.3	13.7	-	9.4	0	10.0	6.3
Operational Readiness	8.9	13.7	-	0	3.3	16.0	0
Sufficient Resources	2.2	2.0	-	0	5.0	0	18.8
Operational Time Factors	2.2	7.8	-	1.9	5.0	4.0	4.2
Personnel & Manning	8.9	7.8	-	1.9	20.0	0	18.8
Timing of Inspections	0	0	14.6	1.9	0	0	0
Attitude	6.7	3.9	-	0	20.0	2.0	10.4
Safety	8.9	17.6	-	0	3.3	6.0	4.2
Adherence to Bureaucratic Stds.	2.2	11.8	-	3.8	5.0	2.0	4.2
Leadership	2.2	0	-	0	11.7	4.0	16.7
Communications Network	4.4	2.0	-	0	0	0	8.3
Physical Environment	0	0	-	0	0	2.0	4.2
Biodemographic Factors	8.9		-	5.7	6.7	0	8.3
Reports	8.9	11.8	-	0	0	10.0	0
Exercises - Drills	-	-	12.2	1.9	-	-	-
Check-off List	-	-	17.1	18.9	-	-	-
Negative (feedback-critical incident)	6.7	9.8	-	0	0	2.0	2.1
Technical Proficiency	0	3.9	-	1.9	10.0	4.0	4.2
Adequate Maintenance	4.4	5.9	-	1.9	5.0	4.0	2.1
Management	4.4	0	-	0	6.7	0	8.3
Evaluate Performance	-	-	22.0	7.5	-	2.0	-
Quality of Equipment	0	3.9	-	1.9	11.7	6.0	6.3

TABLE 3: RESPONSE BY QUESTION

Response	Department Questions			
	Applied			Ideal
	#5	#20	#29	#12
	Eval Method	Eval Criteria	Factors Eff Dept	Methods Should Use
Meeting Operational Commitments	16.7	20.0	5.1	11.8
Efficiency (monetary-manpower)	0	7.5	2.6	5.9
Training	3.7	10.0	20.5	0
Personality (initiative-appearance)	0	0	10.3	0
General Human Relations	0	0	2.6	0
General Evaluation Criteria	7.4	12.5	0	8.8
Operational Readiness	0	7.5	0	0
Sufficient Resources	0	0	2.6	0
Operational Time Factors	1.9	10.0	2.6	0
Personnel & Manning	0	2.5	7.7	2.9
Timing of Inspections	0	0	0	0
Attitude	1.9	0	10.3	0
Safety	0	7.5	0	0
Adherence to Bureaucratic Stds.	5.6	7.5	0	0
Leadership	0	2.5	12.8	11.8
Communications Network	0	0	15.4	0
Physical Environment	14.8	5.0	2.6	0
Biodemographic Factors	0	0	7.7	2.9
Reports	13.0	10.0	0	2.9
Exercises - Drills	0	-	-	0
Check-off List	3.7	-	-	0
Negative (feedback-critical incident)	13.0	12.5	2.6	2.9
Technical Proficiency	0	0	15.4	0
Adequate Maintenance	1.9	2.5	2.6	0
Management	1.9	0	5.1	11.8
Evaluate Performance	1.9	-	-	0
Quality of Equipment	3.7	5.0	2.6	0

TABLE 3: RESPONSE BY QUESTION (Con't)

Response	Goal Related Questions				
	Applied			Ideal	
	#23	#24	#26	#22	#25
	Goals of Unit	Method Assess Goal	Personal Goal to Unit	Goals Should Have	Methods Could Be Used
Meeting Operational Commitments	47.9	20.5	30.8	22.9	37.5
Efficiency (monetary-manpower)	12.5	4.5	1.9	6.3	31.3
Training	16.7	0	15.4	14.6	0
Personality (initiative-appearance)	6.3	0	0	0	0
General Human Relations	2.1	0	0	6.3	0
General Evaluation Criteria	0	2.3	0	0	6.3
Operational Readiness	10.4	2.3	3.8	12.5	6.3
Sufficient Resources	6.3	0	0	2.1	0
Operational Time Factors	2.1	6.8	3.8	2.1	12.5
Personnel & Manning	0	0	0	2.0	0
Timing of Inspections	0	0	0	0	0
Attitude	4.2	2.3	1.9	4.2	6.3
Safety	10.4	2.3	3.8	12.5	6.3
Adherence to Bureaucratic Stds.	2.1	0	3.8	2.1	12.5
Leadership	4.2	0	9.6	2.1	0
Communications Network	2.1	0	7.7	10.4	0
Physical Environment	4.2	2.3	1.9	8.3	0
Biodemographic Factors	4.2	2.3	5.8	0	0
Reports	4.2	4.5	1.9	0	12.5
Exercises - Drills	-	2.3	-	0	12.5
Check-off List	-	4.5	-	-	0
Negative (feedback-critical incident)	0	2.3	0	0	18.8
Technical Proficiency	4.2	0	3.8	2.1	0
Adequate Maintenance	0	2.3	5.8	4.2	0
Management	4.2	0	9.6	0	0
Evaluate Performance	-	11.4	-	-	12.5
Quality of Equipment	8.3	0	0	4.2	0

TABLE 3: RESPONSE BY QUESTION (Con't)

Effect of Organizational Structure on Electronics Maintenance aboard Navy Ships

Herman L. Williams
Lloyd S. Standlee

Navy Personnel Research and Development Center
San Diego

Problem

As a possible solution to the problem of inadequate electronic readiness, the Navy has developed and implemented a new organizational concept aboard a pilot group of medium to large size combatant ships. A comparison of these experimental ships with conventionally organized control ships is required to determine if the reorganization is helping to solve the electronics readiness problem.

Background

Aboard the conventionally organized ship, the responsibilities for electronics maintenance of the major sensor and weapons systems are divided between the Operations and the Weapons Departments. The Operations Department is responsible for maintenance of communications, search and detection, electronic warfare, and tactical data subsystems. The Weapons Department is responsible for maintenance of weapons fire control and delivery, and anti-submarine warfare subsystems. The ratio of unrestricted line to technical officers on these ships is high.

In the pilot organizational structure, the responsibility for all electronic maintenance is assigned to a newly created Combat System Department, and a new billet is provided for system level testing, diagnosis, and fault isolation. The ratio of unrestricted line to technical officers is the reverse of control ships, with most of the officers having had technical education and experience.

Approach

The approach used in comparing maintenance effectiveness aboard pilot and control ships included selection of operational periods, identification of common subsystems, selection of criteria, and statistical analysis of the data.

Selection of Operational Periods

The operating schedules of pilot and control ships could not be standardized. To help minimize differences, operational periods were selected to provide a degree of similarity in type of activity. Periods during which ships were undergoing overhaul were excluded. In any given comparison of a pilot ship with one of its controls, exactly the same months of the year were used. A preliminary analysis failed to show that at-sea time affected the amount of time spent maintaining equipment, so no attempt was made to take this variable into account.

Identification of Common Subsystems

No two pilot and control ships were exactly alike in terms of specific equipment making up the combat system. Therefore, a decision was made to compare maintenance effectiveness on the basis of common subsystems. This was accomplished on the basis of data recorded under the first two digits of the Equipment Identification Code (EIC). The computer search of EICs showed that a pilot ship in many cases shared an individual subsystem with only a single control ship. For example, a pilot ship might have had subsystem A in common only with control ship 1, subsystem B in common only with control ship 2, and subsystem C in common only with control ship 3. To have required commonality of subsystems among all four ships as a basis of comparison would have eliminated subsystems A, B, and C. It was therefore decided to compare each pilot ship separately with each of its control ships, with a total of 12 pilot-control comparisons for each criterion of maintenance effectiveness.

Selection of Criteria

The criteria of maintenance effectiveness were based on the data collected and supplied by the Maintenance Support Office (MSO) 3M reporting system. These data are reported routinely by ships' technicians and entered into a central computer file. The measures relevant to combat systems electronics maintenance, and taken directly from the 3M computer tapes, were:

1. Number of maintenance actions
2. Total days of downtime
3. Total man-hours of maintenance
4. Average downtime per maintenance action
5. Average man-hours per maintenance action
6. Number of personnel-related maintenance accidents.

Separate analyses were performed for repairs made with and without deferral because of lack of spare parts or requirements for outside assistance.

Statistical Analysis

The comparison of maintenance effectiveness, based on common subsystems, is equivalent to an experimental design involving matched-pairs observations. The Wilcoxon Matched-Pairs Signed-Ranks Test was used for those evaluations based on total and average days of downtime and on total and average man-hours of maintenance. The Chi-Square Test was used for those evaluations based on number of maintenance actions and on number of personnel-related maintenance accidents. All tests of significance were at the .05 level of confidence.

Results

Number of Maintenance Actions

Nine control ships reported fewer maintenance actions than pilot ships, two control ships reported more maintenance actions than pilot ships, and in one comparison the difference was not significant.

Downtime and Maintenance Man-Hours

Of the 24 possible comparisons between pilot and control ships, based on a combination of downtime and man-hours of maintenance without deferral for parts or outside assistance, 15 were in favor of control ships, and seven were of no significant difference. Two comparisons were not made because of missing data.

Of the 24 comparisons of downtime and maintenance with deferral for parts or outside assistance, 14 were of no significant difference, four were in favor of pilot ships, and four were in favor of control ships. Two comparisons were not made because of missing data.

Number of Personnel-Related Maintenance Accidents

Five control ships reported fewer personnel-related maintenance accidents than pilot ships. In seven comparisons the differences were not significant.

Summary of Results

Of the 72 comparisons made between pilot and control ships, 29 indicated that there was no significant difference in maintenance effectiveness, 33 favored control ships, six favored pilot ships, and in four cases the data were incomplete. Thus, the findings fail to demonstrate that improved maintenance effectiveness resulted from implementation of the combat system organizational structure on board pilot ships.

Comments

The data are judged to be subject to several sources of experimental error, including:

1. The operating schedules of ships were not standarized.
2. No two ships had exactly the same equipment configuration, and there is a variety of specific equipments within a subsystem as defined by the 3M classification system.
3. There may be considerable variability in ships' failure to report or error in reporting 3M data.
4. The maintenance effectiveness of certain pilot ships may have been confounded by the presence of civilian technical representatives.
5. Insufficient time had elapsed for the pilot organizational structure to have become stabilized.

However, it is believed that the first three sources of experimental error have an equal chance of affecting pilot and control ship data and that the fourth source of error, favoring pilot ships, tends to be counterbalanced by the fifth source of error, favoring control ships.

It must be remembered that the present study concerns itself with a fairly limited definition of maintenance effectiveness. No doubt the pilot organizational structure will have other impacts--for example, on management relationships, career development, morale of crew, and operational performance--that are only partially related to maintenance effectiveness and that merit separate consideration.

Conclusion

The present data do not support the combat system organizational concept. However, the data are somewhat limited and other factors--some of which have been researched and reported separately--should be considered before a final decision is reached concerning implementation of the combat system organizational structure aboard all Navy ships.

Human Performance in Total Ship Planning and Acquisition

Prepared by:

Dr. James McGuinness
ESSEX Corporation

James P. Jenkins
Naval Sea Systems Command

Presented at the Military Testing Association
17th Annual Conference
Indianapolis, Indiana
15-19 September, 1975

INTRODUCTION

This paper will report on the background, current status and future direction of a project designed to develop, test and evaluate an improved Human Factors technology. The project has been designated the Human Factors Engineering Integration Project and is designed to better support Navy Program Managers and Design Engineers in implementing a Total Ship Planning and Acquisition Process, and to demonstrate the basis of this support in quantitative terms.

Human Factors (HF) is broadly defined within this paper as a scientific and engineering discipline concerned with analyses, research, design development and evaluation of the interfaces between man and other men, machines, information and environments. Human Factors Engineering (HFE) is one subclass of HF and includes the engineering aspects of these interfaces. Human Factors in the Navy must be considered in a much broader context than the Human Factors Engineering traditionally used in control or display studies. The issues in the Navy require traditional HFE, but also an increasing emphasis on "macro-level" interactions among such factors as social considerations, economic considerations and personnel test considerations. An example of a basic interaction is that it is difficult to develop criterion for personnel testing on a sonar display without considering whether the equipment is adequately designed. In some cases the social environment must receive as wide a consideration as an equipment factor in a subsystem. An alienated sonar operator has the potential for disturbing an adequately designed system.

Project Objectives

The Human Factors Engineering Integration (HFEI) project is a part of a Navy Advanced Development Plan and has two basic objectives. The primary one is to bring an integrated technology into the ship planning and acquisition process to improve the ship designers understanding and use of data, techniques and methods related to Human Factors. A secondary objective is to bring to the Human Factors field a number of techniques and data that are not normally employed as tools by the Human Factors Engineering specialist. Both objectives are important in any effort to impact the planning and acquisition cycle from a system or platform point of view.

The Advanced Development Plan directs that the HFEI project integrate all HFE disciplines which increase system effectiveness and are cost conscious and effective within the Navy management and design process. A supplemental consideration is to demonstrate integration methodology across Naval Ships Combatant Systems. The driving forces behind this project include the Advanced Development Plan as the primary input. The forces could be further defined as a requirement to more adequately supply HFE data, techniques and methods to designers, especially in advanced technology areas such as the computer-aided ship design system now under development. There is also a need to assess HFE information which is available to Naval ship planning efforts for technical, operational, financial and administrative soundness. This is all focussed around a desire to be responsive to economic, personal and social issues, in addition to the man-machine interface factors that face the Navy Ship Designer today as well as in the future.

Description of the HFEI Project

The David W. Taylor Naval Ships Research and Development Center (DTNSRDC), Annapolis, Maryland is the Technical Agent for the HFEI project. As one of its first implementing actions, center personnel convened a workshop on 18-19 June 1974, and brought together a group of nationally-recognized scientists from the Human Factors Engineering field and from disciplines related to Human Factors Engineering. The results of that workshop stressed the need for a new or revised human factors engineering model, and specified the characteristics of such a model.

Upon completion of this initial planning action, DTNSRDC established a multi-disciplinary project team, composed of several government, university and industrial organizations, as the primary implementing mechanism. The team members and their areas of responsibility were as follows: David W. Taylor Naval Ship Research and Development Center, Annapolis, Maryland (Technical Direction and a description of the Total Ship Planning, Design and Acquisition Process); Naval Electronics Laboratory Center, San Diego, California (Human Engineering): George Washington University (Economics); Boeing Corporation, Seattle, Washington (Computer-Aided Function Allocation and Evaluation System (CAFES); W.D. Teague Associates, New York (Industrial Design); University of Maryland (Social Factors); and Essex Corporation (Integration Contractor).

Essex Corporation was also responsible for an examination of the plans and program for a major Navy system acquisition, the DD 963 (USS SPRUANCE class destroyer). This was accomplished by a review of the Litton proposal and through the experience base of Essex personnel in commercial shipbuilding and Navy programs. The questions of importance in this examination were not those directly related to whether or not commercial shipbuilders were acting in compliance with the ship planning and acquisition process demands and constraints; rather it was oriented toward a description of what had been commercially defined as Human Factors Engineering technology and was being applied in response to these demands and constraints.

Technical Rationale

The technical rationale behind this project hypothesized that there is a body of Human Factors Engineering technology which can be assessed and integrated for application into the total ship planning and acquisition process. Previous Human Factors Engineering technology projects have addressed equipments and/or selected functions of Navy surface ships but not the total ship and the optimization of its full operational capability. It was hypothesized that available Human Factors Engineering technologies are potentially generalizable across platforms (ships, airplanes and submarines) and assoicated classes of problems.

The Components of the First Approach to HFEI

The first approach to a Human Factors Engineering Integrated Technology discussed in this paper is composed of the following elements:

- A preliminary description and functional analysis of the evolving Ship Planning, Design, and Acquisition Process, with emphasis on the Design Phase.
- A proposed structure for assessing and integrating the HFE technologies in terms of scientific/technical, operational and financial criteria; and for relating the integrated technology base to the multi-disciplinary requirements of the Ship Planning and Acquisition Process.
- A review of selected traditional and non-traditional Human Factors Engineering disciplines and technologies, including commercial shipbuilding practices; and an enumeration of technology gaps and needed research.
- A proposed FY 1976-79 project plan for the further development, test and evaluation of the HFEI technology with particular attention to the guidelines and resources required for T&E demonstrations.

The first approach included three technical stages for the development of HFEI. The three stages were conceptualized as descriptive, analytical and integrative. That is:

(1) Descriptive of the available HFE technology and the functional requirements and end-products associated with the total ship planning, design and acquisition process.

(2) Analytical in the assessment of the quality and potential utility of the available technology considering several operational, technical and financial criteria.

(3) Integrative in the sense that a revised structure of HFE was initiated which includes both traditional and non-traditional HFE disciplines, which directly supports total ship planning and acquisition process requirements, and which includes a proposed approach for combining HFEI technologies. When fully developed, a new operational capability will permit combinations of accessed and catalogued HFEI technologies to be selected to fill the administrative and technical requirements of specific ship development programs.

The Ship Planning Design and Acquisition Process

The Ship Planning, Design and Acquisition Process was described as generally consisting of the following three major stages:

(1) "Need" identification or, requirement derivation stage

(2) Design stage

(3) Production stage (Detail Design and Construction Stages)

In support of the process, major studies are directed by the Chief of Naval Operations to develop an Operational Requirement (OR) document, which states an understanding as to the ship type and approximate size desired. At this point, analyses and studies are concerned with Mission and Force Leval Trade-offs. Then feasibility studies are conducted to examine system level trade-offs such as payload versus protective needs. A ship platform configuration is selected during the concept design stage.

The Design stage goes from concept design through preliminary design to contract design. The Production stage, with the major components of detail design and construction, was described and specified in the HFEI project as were the subsequent T&E and Operational Evaluation (OPEVAL) procedures.

Major impact points for Defense System Acquisition Review Council (DSARC) decision points, as well as other trade-off and selection processes were defined in greater detail in the first approach to HFEI. The detailed coverage examined design events and decisions during each stage and the implications for HFEI, as well as required HFEI inputs. These descriptions were necessary to lay a foundation for understanding the ship planning process in order to understand more fully how past HFE efforts impacted Navy ship systems, as well as to determine how future efforts could be more effectively incorporated into the system design process.

Proposed Structure for Assessing and Integrating HFE Technologies

The first approach to the HFEI brought together a number of technologies which can be effectively integrated into the Ship Planning and Acquisition Process. This approach examined areas of technology which have been used by Human Factors Engineers but also presented areas which have not often been previously coordinated with Human Factors activities (e.g. social factors and economics).

The five areas examined in the first approach were:

(1) Human Engineering
(2) Computer Aided Function-Allocation and Evaluation System (CAFES)
(3) Industrial Design
(4) Social Research Applications
(5) Economics

These technologies were examined in some detail by the other HFEI project participants under contract to DTNSRDC. Essex Corporation, as part of the integration effort, abstracted the major details from each report and classified the theories, techniques, methods, data and bibliographies for further assessment and incorporation into an HFEI technology.

Table 1 demonstrates one type of assessment procedure suggested for reporting on a specific subject in any literature evaluation. Table 2 contains a preliminary assessment and scaling methodology developed to evaluate individual technologies, methods or theories. Both of these techniques will be modified on the basis of empirical validation experiences. They do indicate the type of technology criteria that will be a necessity for the analytical and integration phases of the HFEI effort.

Future Directions

One of the major emphases in this project could be entitled "relatedness" as well as integration. The project must relate to Navy processes such as the Ship Planning and Acquisition Process and to other pertinent processes such as the T & E process. T & E problem areas regarding HFEI in ship systems will have to address major questions such as the effectiveness in total ship systems versus effectiveness in subsystems such as Command, Control and Communications, Power Management, Weapon Systems, Life support etc..

To address these processes, criteria development and application for HFEI must be subjected to feasibility testing involving three practical levels. They are 1) Operational Feasibility, 2) Technical Feasibility and 3) Financial/Administrative Feasibility. The following kinds of questions will be applied to feasibility testing of the criteria:

Operational feasibility test criteria need to consider questions such as the following:

(a) How adequately does the HFEI technology support System Planning and Acquisition ?

	Source #1	Source #2
1. Title	Display Evaluation Index (DEI)	Computerized Link Analysis
2. Description	Calculational method for deriving a figure of merit of display effectiveness in transferring to and from the operator	Technique to arrange men and equipment based on frequency sequence and criticality of use
3. Source	Applied Psychological Services	Autonetics
4. Reference	Human Factors, 6/64, v. 6, no. 3, pp. 279-286	Human Factors, 2/64, v. 6, no. 1, pp. 63-70
5. State of Development	completed and tested satisfactorily	completed but probably obsolete
6. Application	any visual display unlimited platform	unlimited platform
7. Interrelationship with other methods	possibly work to advantage in a computerized human engineering system	highly dependent upon having inputs from OSD's
8. Quantification	Highly	subjective ratings-i.e. relative importance certain links
9. Cost	could complete fairly complex design in two days	50% dollar/67% time savings (autonetics data)
10. Advantages /benefits	relatively simple, quantitative, valid and reliable technique for assessing display merit	handles large systems beyond manual capability
11. Problems /limitations	requires some algebraic sophistication	number-coding technique cannot handle more than 28 elements per console

TABLE 1 EVALUATION METHODOLOGY

	SCALING OF TECHNOLOGY*					
	-2	-1	NO RESPONSE NR=0	+1	+2	RESULTANT CODE
STAGE OF DEVELOPMENT						
A - Conceptual B - Evolving, Modifications Required C - Final Testing Required D - Completed	A	B	NR	C	D	
					X	+2
ASSUMPTIONS REQUIRED	Many	Some	NR	Few	Very Few	
- Definition of Behavioral Criteria Important				X		+1
CONDITIONS FOR USE	Many	Some	NR	Few	Very Few	
- Degree of Practitioner Skill and Experience					X	+2
APPLICATION "COSTS"	Very High	High	NR	Low	Very Low	
- ROI (Time/cost) Depends on Systems Complexity				X		+1

* AN EXAMPLE OF THE TYPE OF SCALING METHODOLOGY PLANNED FOR THE ASSESSMENT PROCEDURE, I.E., A PLUS (+) 2 INDICATES A VERY FAVORABLE VIEW OF THE TECHNOLOGY A MINUS (-) 2, INDICATES A VERY UNFAVORABLE VIEW OF THE TECHNOLOGY FOR THE CRITERION.

ASSESSMENT AND SCALING METHOD APPLIED TO OPERATIONAL SEQUENCE DIAGRAMS

TABLE 2

	Very Low	Low	NR	High	Very High	
ACCESSIBILITY - Computer vs. Manual Computerized OSD Available - Guidebooks - Manuals - Classified, Proprietary, Available					X	+2
DATA BASE - System Applications (Models etc.) - Statistical Analysis					X	+2
VALIDITY - Overall - If Conducted Properly, Very High - Content, Predictive, Construct					X	+2
RELIABILITY - Tools, Techniques, Methods					X	+2
GENERALIZABILITY					X	+2
USEFULNESS - Can Establish Training Objectives					X	+2

TABLE 2 CONTINUED

(b) Is the HFEI technology clearly responsive to Program Managers' and Project Engineers' requirements?

(c) How well is the HFE need matched by the new HFEI technology?

(d) How much does the application of HFEI technology improve ship material, manpower and facilities readiness and operational effectiveness?

(e) Will the application of HFEI technology help meet mission, design and other objective, e.g., reduce manpower and training requirements?

(f) To what degree does the new HFEI technology help resolve the often conflicting criteria of costs, system performance, human performance, team structure, etc.?

Technical feasibility test criteria need to consider questions such as the following:

(a) How much improvement in the quality of HFE procedures, data and techniques has been achieved in the HFEI technology? Can critical multi-disciplinary problems be dealt with effectively?

(b) What new/improved technical solutions can be demonstrated? Against what criteria?

(c) What specific improvement in tradeoff capabilities (e.g., design and human performance) have been achieved?

(d) What operational improvement could result from the technical advancements?

(e) To what degree can the HFEI technology data system be verified?

(f) What gaps in man-machine-environment interaction data exist? Can they be filled by broad application of HFEI technology?

Financial feasibility test criteria need to consider questions such as the following:

(a) Will increased costs for systems planning and acquisition result from application of HFEI technology? Are the increased costs justified by return on investment?

(b) Will the implementation of HFEI technology require more dollars? manpower? facilities, including computer support?

(c) Will HFE personnel require extensive training to implement the new technology? At what cost?

(d) Will the application of the HFEI technology extend the time required for administrating the ship planning and acquisition process? At what cost?

(e) What additional RDT&E funding is required to fill the major technology gaps?

The answers to the above questions will bring a practical realization to the HFEI technology. It will also begin to define system test loci and methods of measurement vs. the ship system parameters of interest. Before this is accomplished, a number of technical challenges regarding performance testing are obvious. The challenges are germane across many sectors of the personnel testing community. They are represented in the following statements regarding technology gaps selected from an initial look at issues involved in developing the HFEI technology:

- Development of a system of data banks for HF data and techniques is needed.
- Design handbooks dealing with anthropometry, weight carrying limits, etc., need to be revised to include more dynamically based data, e.g., the projectile weight limits for ammunitions handling on a ship over a sustained performance period under realistic conditions.
- The effects of various sea state conditions on human performance are not well understood or documented.
- The interaction of manning and economic factors should be formulated as they effect the design process and life cycle costs.
- Survey methodology needs to be examined for its validity and reliability as applied to Navy social problems. Pre-administrative pilot tests of questionnaire items and techniques must be required more routinely.
- Experimental designs, analyses and field testing should more often employ multi-variate analysis where appropriate.
- Multi-dimensional scaling techniques should be used with more regard to the validity and reliability considerations involved.
- Increased emphasis should be place on the Validity and Reliability bases for models and techniques.

- Modeling must consider multi-measurement approaches, e.g., time measures of an operator station must consider associated error rates;tracking scores must consider accuracy as well as time, etc..

Summary of the Next Stage in HFEI Development

Briefly summarized, the FY 1976-79 proposed program plan will provide for accomplishing the following:

- Expanding the scope of the first approach by
  - (a) more in-depth examination and detailed description of the Total Ship Planning and Acquisition Process
  - (b) an increase in the number of HFE technologies reviewed and assessed, both traditional and non-traditional
  - (c) further development of the assessment and integration process with emphasis on the ship parameters and classes of problems as the basis for the HFEI taxonomy/matrix
  - (d) further examination of commercial shipbuilding HFE practices by site visits and interviews
- Consolidation of all inputs
- Specification of test and evaluation plans, resources, designs and criteria

Concluding Remarks

The ultimate objective of the HFEI technology program is to produce an operational capability, probably some combination of computer and non-computer modes, in which the state-of-the-art in Human Factors Engineering has been comprehensively described, assessed and integrated so that Program Manager/Project Engineers and Human Factors Engineering Specialists can quickly retrieve and apply whatever combinations of technologies meet their scientific/technical and administrative criteria for particular ship system developments.

The HFEI program must be designed and redesigned as necessary with this ultimate objective in mind.

Lessons Learned From the First Approach to HFEI

- Large System HFE integration technology requires definitions of input technologies before "integration" can be developed.

- Objective and standard assessment methodology must be developed to effectively integrate HFE techniques.

- Economic theory does provide technical methods which can be assimilated into HFEI (e.g. production functions).

- Methodologies resident in social study areas can be used for ship HFE applications (e.g., need assessment technique).

- The technology base in industrial design is difficult to quantify within the ship design process, although the results may be worth the effort in particular instances.

- Models must incorporate subsystem integration to be useful to system designers and analysts.

- The Ship Design Process is very iterative, and in practice can quickly generate a frozen design. The process is evolving and is becoming computer based. HFE is not a major contributor to the technology base in practice, or in the planned computer base.

- Commercial HFE practices in ship design are advisory, based on generalizations and recommendations, are easily modified and/or neglected in-house or by the customer.

A Bayesian Method for Maximizing Correct Mastery Classifications

Frederick Steinheiser, Jr., PhD
Army Research Institute for the
Behavioral and Social Sciences
Arlington, Virginia 22209

Introduction

This paper summarizes work which is part of an on-going research program at the Army Research Institute. The program, called Mettest--Methodological Issues in the construction of Criterion-Referenced Tests--is exploring and developing psychometric models for defining test standards and test lengths. The focus of this paper is upon a "Bayesian" model, which gives the probability of correctly classifying an examinee as a master or as a nonmaster while taking into consideration the test length and the mastery cut-off score.

Personnel assessment is typically made through the development and administration of tests, and the evaluation of test scores. The final desired output of a test for a given examinee is information which allows us to pinpoint his ability to do whatever is required by an objective. That is, we observe a test score and must then infer the ability of the examinee.

In Norm-Referenced Testing, an examinee's score is evaluated with respect to his position among all of the other scores in the examinee population. But in Criterion-Referenced Testing his score is evaluated with respect to his passing or not passing a particular instructional objective, independent of the scores obtained by others in the examinee population. A passing score indicates that he is a "master" of that particular instructional objective, and a failing score indicates that he is a "nonmaster."

Ideally, if an examinee's score on a test is above the minimal passing standard, he would be correctly classified as having "mastery ability." Ability assessment would therefore be based upon 100% correct classifications. But since we live in a less than ideal world, there will be variability due to the imperfections in the test construction, psychological variability (forgetting, guessing, individual differences) in the examinee population, and other unknown sources of error. Hence, sometimes a person who is really not a master (in the ideal world) will be classified as a master on the basis of his test score; and sometimes a person who really is a master will be unfortunate enough to be classified as a nonmaster. The following chart illustrates these four classification outcomes.

The views expressed in this paper are those of the author and do not necessarily imply endorsement by the U.S. Army.

		True Ability State:	
		Master	Nonmaster
Classification Based Upon Test Score	Master	True positive	False positive
	Nonmaster	False negative	True negative

In order to approach the ideal classification accuracy, the probability of a True positive should be much greater than that for a False positive, and the probability for a True negative should be much greater than that for a False negative. The classification problem has now been cast into a decision-making framework, for which "Bayes' Theorem" may be used: we wish to obtain the probability of an examinee being in the Mastery Ability state given (conditional upon) his test score. Symbolically, this is expressed as $p(M1|T)$, where M1 refers to the mastery state, and T is the test score of that specific individual.

An Example Using Bayes' Theorem

Bayes' Theorem is a mathematical expression which allows us to combine information about the quality of the examinee population so as to produce a probabilistic estimate of mastery for a specific examinee. This approach will give the most accurate ability estimate for each examinee by using the fewest number of test items, provided that accurate estimates of the "quality parameters" have been made. (A subsequent section of this paper will show exactly how to estimate these quality parameters from data similar to that which you might have.) First, let's take a look at the mathematical expression, the estimates that we have to feed into it, and the output that it gives us:

$$p(M1|T) = \frac{p(T|M1)p(M1)}{[p(T|M1)p(M1) + p(T|M2)p(M2)]}$$

Here we assume that the 2 states of nature (master and nonmaster) are mutually exclusive and collectively exhaustive, and that T is the test score which is observed. We also assume that the test is dichotomously scored. A correct response is denoted "1", an incorrect response is denoted "0" and the total test score is simply the number of correct responses. What we seek to find is the term on the left, the probability that a given student is a master, having been given his test score. In order to find it, we need to have an estimate of the prior probability of mastery (p(M1)) in the population of students from which this student was drawn. The prior probability of mastery can be thought of as the proportion of students in the examinee population we think are masters. For example, if our instruction were very good the prior probability of mastery would be high, and most of the students who completed the instruction should have mastered the objective. The actual number specified for the prior probability of mastery may be an informed guess based on experience or it may be based on the empirical results of tests given to previous classes of similar students.

We must also estimate the conditional probability of a certain test score given that the student who got that socre was a master. For example, if only one item were administered, the conditional probability of a score of one correct given that the student was a master is simply the probability that a master responds correctly. We may estimate this conditional probability empirically based on previous student groups, or we may provide a best guess as to how well masters perform, or this conditional probability may reflect a minimal standard of achievement. We shall show how the $p(M|T)$ will vary as a function of the prior expectations of the tester, number of test items, and conditional probabilities, $p(T|M)$, after an example to illustrate the computations.

Suppose that a student chosen at random from a trainee population was given a criterion-referenced test, and that he passed the test. Given the results of the test, what is the probability that the student is indeed a master of that particular course of instruction? In order to calculate the probability, we obtain the following information from the educational expert who administered the CRT: The probability that a master would obtain a passing score = .90, ($p(T|M1) = .90$); the probability that a nonmaster would obtain a passing score = .05, ($p(T|M2) = .05$); and the prior probability of randomly selecting a master from this trainee population is equal to .70, that is, we believe that 70% of this and similar previous trainee populations may be assumed to be composed of masters. Substituting these values into the formula;

$$p(M1|T) = \frac{.9 \times .7}{.9 \times .7 + .05 \times .3}$$

which equals .977. Hence, before the test score was available, the probability that this student was a master was .70, but after a passing score was observed, the probability that this person is a master has increased to .977. (The probability of this student's being a nonmaster, given the same passing score, $p(M2|T)$, would be equal to 1-.977 or .023).

Methods for Estimating "Quality Parameters" (Prior and Conditional Probabilities)

This model assumes that background information about students who previously took the test is available. This background information should lead to the accurate estimation of parameters that describe the quality of the examinee population. We need information to estimate the prior probability of a randomly selected student being in one of the assumed mastery ability groups. We also need to be able to estimate the conditional probability that an examinee from a particular ability group would get an item right. For purposes of illustration, let's assume that 100 examinees produced the following distribution of scores on a five item test:

Number of items correct	Frequency	p(Correct)
5	30	1.0
4	30	.8
3	00	.6
2	10	.4
1	20	.2
0	10	0.0

For this particular set of data, it seems reasonable to postulate two ability groups. Note that 60 examinees got either 4 or 5 test items correct, and that a total of 40 examinees got 0, 1, or 2 items correct. No one got 3 items correct. Hence, this bimodal distribution of scores strongly suggests that we may set the prior probability of mastery equal to .6, and the prior probability of nonmastery equal to .4. Symbolically, $p(M1) = .6$, and $p(M2) = .4$.

We also need to estimate the conditional probability that a correct response is made to an item, given the particular mastery or ability state. Symbolically, we seek $p(x=1|M)$. There are several ways to compute this parameter value. Unfortunately, each method produces a unique value.

We could take the average proportion correct for the mastery group. For the present data, this would produce $\frac{30 \times 1.0 + 30 + .8}{60} = .9$.

We could also take the lowest score in the mastery group, which in this case is .8.

We could also take the desired (or "standard") score required for the demonstration of mastery, which need not necessarily be observed. This value for the present data could be set at .70, or .71, or .82, etc.

This variety of estimated values should not be distressing, since it allows the examiner to introduce his own requirements into the selection. The important thing is that his choice should be close to at least one of the logically derived estimates.

The Bayesian Mathematical Model

In order to generalize the Bayesian approach to a wide variety of applications in personnel assessment, two additions must be made to the previously described formula. These additions are the number of trials or items on the test (N), and the number of hypothesized mastery ability states (S). The derivation of the general formula to meet these goals was originally presented by Hershman (1971):

$$p(Mi|T) = \frac{\prod_{j=1}^{N} p(Mi|t_j)}{p(Mi)^{S-1} \sum_{i=1}^{S} \frac{\prod_{j=1}^{N} p(Mi/t_j)}{p(Mi)^{S-1}}}$$

In this formula, $p(Mi|tj)$ equals the conditional probability of a person in the ith mastery state getting the jth test item correct; p(Mi) is the prior probability of the representation of the ith mastery state in the student population (the % of students who are estimated to be in the ith mastery state); and $p(Mi/T)$ is the conditional probability of a particular student being classified as being in the ith mastery state given his total test score. A computational example showing how the formula is applied for three mastery states is given in the Appendix.

This formula was not used to analyze any "real" test score data. Rather, selected values of the various parameters were systematically manipulated in order to determine their influence on the probability of mastery classification. Hence, the results are from a computer simulation of idealized data and should serve to emphasize the relative effects of each of the parameters.

The two parameters that estimate the quality of the examinee population, the prior probability of selecting a master from the population, and the conditional probability of a known master and of a known non-master getting a randomly selected item correct, have already been discussed. Basically, p(M1) reflects the proportion of masters in a group of examinees (and the level of training), whereas $p(1|M1)$ and $p(1|M2)$ will be high if the test is easy and lower if the test is difficult. The ideal criterion referenced test should provide a high probability for the former, and low probability for the latter.

Two other parameters, the minimal passing standard (the per cent of all the items that were answered correctly) and the test length are interrelated. By way of analogy, consider the minimal passing standard for deciding that a coin is biased to be 70%. That is, if heads (or tails, for that matter) come up on 70% of the tosses, we would evaluate the coin as being unfair. Note that the 70% figure is arbitrary. We could have set the standard at 65%, or 75%, or 80%, etc. Now how many tosses (test items) do we want to observe? 10? 50? 100? 1,000? If we observe 7 heads out of ten tosses, and 700 heads out of 1,000 tosses, does the probability of the fairness of the coin remain the same? It should be intuitively obvious (and it can be easily demonstrated by means of the binomial distribution) that the minimal passing standard interacts with test length. The probability that a coin is fair when 7 heads out of ten tosses are observed is much greater than when 700 heads out of 1,000 tosses are observed--even though the "70%" criterion was strictly maintained! Values of 60%, 70%, and 80% correct were used in the current simulation. The number of trials or items (N) took on values of 5, 10, 20, and 40.

The final parameter which was manipulated in the model is the number of assumed mastery states. It may be overly simplistic to assume that the world is divided into just two dichotomous and mutually exclusive states, of mastery and nonmastery. Perhaps there are varying degrees of mastery, ranging from "complete" to "partial" to "totally incomplete." The present model is able to handle any number of assumed mastery states.

The model makes the following two important assumptions concerning the nature of the test from which the data are derived: (1) The test measures a unidimensional latent trait or unitary skill; (2) Test items or trials are equally difficult for a given ability. An elaboration of the basic model can easily be made to include test items with varying degrees of difficulty.

Changes in p(M|T) Assuming Two Mastery States

The fundamental purpose of the present study was to investigate how the probability of mastery classification changes as a function of the simultaneous manipulation of up to four parameters (independent variables). The scope of the study is not exhaustive, since only several were used. However, some general trends do seem to emerge as can be seen in the following figures.

Figures 1, 2, and 3 show the results of applying the model to a situation in which only two mastery groups (mastery and nonmastery) have been hypothesized. The data points represent the probability that a trainee is a master, given (conditional upon) his total test score, P(M|T). The curvature of each line shows how the P(M|T) changes as a function of variations in the prior expectation of mastery, the % correct items observed, the conditional probabilities of both a master and a nonmaster responding correctly to an item, and the number of items comprising the test.

Figure 1 represents a testing situation in which the training was of extremely high quality, since the proportion of masters in the trainee population was assumed to equal 0.9. That is, p(M1) = 0.9. Figure 1A portrays the situation in which both masters and nonmasters have attained a rather high degree of proficiency, since the probability of a master responding correctly to any given item is 0.9, and the probability of a nonmaster responding correctly is 0.6. If a person scored 80% on a five item test, the probability that he is a master is approximately .91. This probability drops to .65 if a 60% score on five items (3 out of 5 correct) were obtained. Note that when the test length is increased to 40 items, an 80% score (32 correct) produces a .99 probability of mastery. However, a score of 60% (24 correct) yields an essentially zero probability of mastery. The effect of the test length variable on classification accuracy is dramatic: if the p(M|T) had to be at least 0.5 for a person to be called a master, then scores of 60% on a five item test would lead to mastery classification. But a 60% score on a 40 item test would lead to nonmastery classification.

Figure 1A also illustrates the effect of "prior beliefs" on p(M|T). Intuitively, one might suppose that the chances were much higher that a person who obtained a score of 60% (even from a 5 item test) came from a population whose probability of correctly answering an item was 0.6 than from a population whose probability of answering an item correctly was 0.9. However, the relative proportions of the two groups (expressed as prior belief in mastery and nonmastery, or p(M1) = .9 and p(M2) = .1, respectively) are such that the probability of a person being in the mastery state is approximately 0.65 for a score of 3 correct (60%) on a 5 item test. Only by increasing the number of test items can the strong prior bias in favor of the mastery decision be reversed. Figures 2A and 3A show what happens when prior beliefs are not so heavily biased in favor of mastery. In neither case is the probability of being in the mastery state above 0.5 for scores of less than 80%. But Figure 1A suggests that when prior beliefs heavily favor one group over the other, longer length tests should be used. Otherwise, the amount of data may not be sufficient to force a change in the originally held prior beliefs.

The effect of changing the prior beliefs concerning the proportion of masters and nonmasters in the examinee population while holding all other parameters constant can be seen by comparing corresponding graphs A, B, C, and D in Figures 1, 2, and 3. As the prior beliefs approach equiprobability (where p(M1) = p(M2) = 0.5), more items are required to maintain a given level of confidence that a person is either a master or nonmaster. The inability to postulate strong prior beliefs must be compensated for by increasing the test length in order to maintain a constant classification accuracy.

The effect of changing the probability of a correct response, p(1|Mi), can be seen by comparing graphs A, B, C, and D for Figures 1, 2, and 3. For example, the only difference between Figure 1A and Figure 1B is that the $p(1|M1)$ changes from 0.9 to 0.8, all other parameters being held constant. (This change might reflect a lower level of required proficiency, and hence less training, for Graph B than for A. Or perhaps previous test results indicate that masters of the instruction respond to items with a probability of correct response equal to 0.8 rather than 0.9.) In any case, the effect of this small change in the $p(1|M1)$ on the $p(M|T)$ is readily apparent. For any test length or observed test score, the probability of being in the mastery state is greater in Graph B than in A. This shift is most obvious for the 70% observed correct curve. Notice that $p(M|T)$ on Graph A for an observed score of 70% (28 out of 40 correct) is approximately *0.04*. However, the value for $p(M|T)$ in Graph B for 70% of a 40 item test correct is *0.87*.

The main reason for this abrupt change from Graph A to B (in Figures 1, 2, and 3) is the lowered requirement for mastery, from 0.9 to 0.8. The probability that "0.9 persons" score only 70% correct on long tests is relatively low. But when masters are defined as those trainees who come from a population with a probability of responding correctly equal to 0.8, the probability of their scoring 70% on a long test is high. One of the most difficult jobs for an instructional designer is to describe the level of capability *required* of graduates and the level of capability actually *achieved*. Comparison of these graphs indicates the magnitude of the effect that these specifications can have on the classification of trainees.

Graphs C and D of Figures 1, 2, and 3 further illustrate the effect of variations in the probability of correct responses. The only difference between Graphs B and C is that the probability of a correct response from a nonmaster decreases from 0.6 to 0.5. The effect of this decrease in correct response probability from a nonmaster is to lower the likelihood of a nonmaster achieving a test score of at least 70%, which also increases the probability that a person achieving a high % score is in the mastery state. Finally, Graph D portrays an extreme case in which neither masters nor nonmasters are responding at particularly high levels. However, the level of performance for nonmasters is so low (0.4), that even for observed scores of 60% the probability of being in the mastery state exceeds 0.8 for all test lengths, except for 5 and 10 items in Figure 2, and 5, 10, and 20 items in Figure 3.

Further detailed analysis of these figures is not included in this paper. In comparing the twelve graphs against each other, note the magnitude of the changes in $p(M|T)$ when small changes have been made in the prior beliefs, in the correct response probabilities, and in the percent correct observed responses. The implication is that extreme care must be taken when specifying parameters in a Bayesian approach to testing and decision making. If the parameters are realistic, great savings in testing time and expence, and increased confidence in decision making are possible (Novick & Lewis, 1974). However, if the parameters are not realistic, there is a very real danger of misclassifying many examinees. The next section of this paper deals with an elaboration of the model to three mastery states, thus helping to quantify sources of classification error.

Elaboration to Three Mastery States

Figures 4, 5, 6, and 7 represent cases for which three mastery states have been hypothesized. In figures 4 and 6 the probability of a correct response for a person assumed to be in mastery state M1 equals 0.8, for mastery state M2 this probability is 0.6, and for mastery state M3 it is 0.5. These values could correspond to the situation in which the non-mastery group was divided in half. That is, those persons whose probability of getting any given item correct is 0.5 (comprising mastery state M3) would need extensive retraining; whereas those whose probability of getting any given item correct is 0.5 (comprising mastery state M3) would need extensive retraining; whereas those whose probability is 0.6 (comprising mastery state M2) would merely need selective retraining. People in mastery state M1 have a probability of 0.8 for making a correct response, and may therefore be considered as "masters" who have successfully passed training.

For Figures 5 and 7 the corresponding probabilities of a correct response for people in mastery states M1, M2 and M3 are 0.9, 0.8, and 0.6, respectively. These probabilities might describe a situation in which the mastery group was dichotomized, perhaps in an attempt to identify those students who had achieved an exceptionally high level of proficiency, i.e., $p(1|M1) = 0.9$.

In Figures 4 and 5 the prior probability (or assumed proportion) of examinees in each mastery state are: $p(M1) = 0.5$, $p(M2) = 0.3$, and $p(M3) = 0.2$. In Figures 6 and 7 the corresponding prior probabilities are 0.25, 0.50, and 0.25, respectively. The prior values in Figures 4 and 5 display a bias towards higher levels of mastery (50% of the examinees are assumed to be type M1 masters), whereas the bias in Figures 6 and 7 is towards the intermediate level of mastery (50% of the examinees are assumed to be type M2 masters).

A detailed analysis of Figures 4 and 5 will provide the basis for an interpretation of Figures 6 and 7, which is an exercise left to the reader. The three graphs labeled A, B, and C represent the probability that an individual is in mastery state M1, M2, and M3, respectively.

Graph A shows the probability that an individual is in mastery state M1 given observed scores of 60%, 70%, and 80% correct on 5, 10, 20, and 40 item tests. Thus, for an observed score of 4 out of 5 correct, the probability that this person is in mastery state M1 is about 0.65. But if this same person got a score of 32 out of 40 (still 80% correct), the probability that he is an M1 master jumps to 0.98. These results are similar to those obtained when two mastery groups were hypothesized, and again illustrate the effect of increasing test length on the level of confidence in the mastery classification $p(M|T)$.

The probability of being in mastery state M2 given observed scores is plotted in Graph B. If a person got 4 out of 5 correct, the probability of being in state M2 is about 0.25. However, if he got 32 out of 40 correct (still 80% correct), this probability plummets to 0.02. Finally, using these same test score values, Graph C shows that the probability of being a type M3 master is 0.10 for 4 out of 5 correct, and nearly zero for 32 out of 40 correct. This result makes intuitive sense, because there is only 20% of type M3 (non)masters in the examinee population, and the probability of their getting any item correct is only 0.50, which is a long way from 80% observed correct.

Notice that for any given test length and percent correct, the sum of the probabilities of being in states M1, M2, and M3 equals 1.0. Comparison of Graphs A, B, and C shows that when either 70% or 80% of the items for any test length are correctly answered, the probability of being in state M1 is greater than the probability of being in either state M2 or M3. That is, both the 70% and 80% curves are higher in Graph A than in either Graph B or C. For an observed score of 60% the probability of being in state M2 is greater than for M1 or M3. The probability of being in state M3 is rather low for all values of test length and percent correct observed in this particular example.

In Figure 5 the interrelationship between test length and three hypothesized mastery states becomes even more apparent. For example, Graph A shows that the probability of being in state M1 for 80% correct on a 5 item test is about 0.48. The probability of being in state M2 (shown in Graph B) for 80% correct on a five item test is about 0.36. There is thus a greater chance that a person whose score is 4 out of 5 is in M1 ($p(M1|T) = 0.48$), instead of M2 ($p(M2|T) = 0.36$) or M3 ($p(M3|T) = 0.16$). However, if a score of 80% correct were observed on a 40 item test, the graphs indicate that a much different decision would be appropriate. In this case, $p(M1|T)$ equals 0.21, $p(M2|T) = .78$, and $p(M3|T) = 0.01$. Hence, people scoring 32 out of 40 correct should be classified as type M2 masters. Also note that a score of 60% for any test length implies that these people should be placed in the M3 state.

For the data used in Figure 5, the probability of finding M1 type masters is overall quite low. Instead, for the levels of achievement demonstrated by obtained scores of 60%, 70%, or 80,it is more likely that such scores were produced by people in mastery states M2 ($p(1|M2) = 0.8$) and M3 ($p(1|M3) = 0.6$).

Test Length and Misclassification Error

One of the most important questions that must be answered in designing a training evaluation program is: What is the probability of falsely classifying a person on the basis of a given observed score? It is also possible to turn the question around and ask: How long must a test be, and what score is required for classification decisions to be made with some specified lower limit of misclassification?

Figures 8 and 9 demonstrate how the Bayesian model can be used to answer the above questions. Assuming that the prior and conditional probabilities are realistic and fixed, the important variables are then test length and cutting score. Suppose that $p(M1) = 0.9$, $p(M2) = 0.1$, $p(1|M1) = 0.9$, and $p(1|M2) =$ [illegible] as in Figure 8. In this example, the prior belief that an untested trainee is a master is very high, $p(M1) = 0.9$. A reasonable question might therefore be: What score must be observed such that a nonmastery decision can be made with at least 90% confidence? In other words, what data are required to force a reversal in the prior belief?

To be 90% confident of a nonmastery decision, $p(M2|T)$ must be equal to at least 0.90. Since the sum of $p(M1|T)$ and $p(M2|T)$ equals 1.0, $p(M1|T)$ must therefore not be greater than 0.10. Referring to Figure 8, a horizontal line crossing the ordinate at 0.10 can be drawn. This line crosses the curve for a five item test at a point corresponding to 26% correct. The next lowest possible test score is one correct (20%), so the decision rule is that all persons scoring one correct or less should be considered nonmasters. The point on the ordinate corresponding to 20% correct on the five item test is about 0.05. Hence, the final decision rule states that nonmastery decisions based on an observed score of one correct out of five can be made with 95% confidence ($1.00 - 0.05 = 0.95$). For observed scores lower than the cutoff score the confidence in making a correct decision must increase. Continuing with the present example, the $(pM1|T)$ if zero correct are observed is virtually equal to zero. Hence, those persons who get no items right may be classified as M2 type nonmasters with nearly 100% confidence.

A similar analysis applied to the 40 item test curve indicates that the cutting score should be about 73% correct. The next lowest possible score to 73% is 70%, which yields exactly 28 correct out of 40 items. The probability of mastery given an observed score of 28 correct is about 0.04. At such a low value of $p(M1|T)$ the chances for misclassification using a five item test and a 40 item test are almost the same. However, the observed percent correct at which the nonmastery decision is made for the two tests is 20% on the five item test and 70% on the 40 item test. Superficially, two tests of different lengths would seem to produce the same decision outcome, and that longer tests may not really be necessary for reducing classification error.

In order to appreciate the benefits gained by using longer length tests, the entire curve must be examined. Note that at 80% correct the five item test yields p(M1/T) to equal 0.92. This result suggests that, on the average, 8% of the mastery decisions will be in error. For the 40 item test, the probability of mastery given 80% correct is about 0.99. That is, there is only about 1% chance of misclassification error. A test that distinguishes sharply between masters and nonmasters is one in which the probability of mastery is close to either 0.0 or 1.00 for most obtained scores. On such tests there is only a small region in which classification error is large. For example, in Figure 8 for the 40 item test, the region where p(M1/T) is greater than 0.1 and less than 0.9 extends from 71% to 77% correct. This means that the probability of misclassifying a person will exceed 0.10 only when observed scores range from 71% to 77% correct. In contrast, the region of the five item test curve for which p(M1/T) is greater than 0.10 and less than 0.9 extends from about 26% to about 79%. Hence, there is a much larger region of the curve for which the probability of misclassification exceeds 0.10. Obviously, if classification accuracy is to be maximized over the entire range of possible test scores, then longer tests are required. Ideally, a very long test would produce a step function, for which all values of possible scores approach either 0.0 or 1.0.

Figure 9 can be analyzed in a manner similar to that for Figure 8. However, Figure 9 has one outstanding characteristic that merits special attention. If nonmastery decisions must be made with 90% confidence, and a horizontal line at p(M/T) = 0.1 is drawn, the line does not intersect the curve for the five item test. This means that it is not possible to classify a nonmaster with 90% confidence if a five item test is used, given the parameters used in Figure 9. If resource or time constraints are such that no more than five items may be given, and if the parameter values used in Figure 9 are realistic, and if 90% confidence for mastery decisions are required, then there is no reason to test. Testing is irrelevant because no matter what score is observed, including zero correct, the decision rule compels a mastery decision to be made. In fact, for the present values, the probability of mastery given zero correct, is equal to 0.21. This simply means that if persons obtaining a score of zero are classified as nonmasters, 21% of them will be misclassified, on the average.

Conclusions

The implications of the results from this simulation experiment stress the practical importance of test length, criterion scores, and accurate estimates of examinee quality in making optimal mastery classifications. If the assumptions of the model have been met, and if accurate parameter estimates have been made, then a Bayesian method is optimal in the sense of providing more accurate estimates of mastery classification with the least number of test items.

In the typical situation for personnel assessment, the examiner will have some degree of control over the values of the parameters. His estimates of the *prior* probability of mastery will depend upon the goodness of the information he can obtain about previous examinee populations' scores. His estimates of the *conditional* probabilities can be made by several equally justifiable and logical procedures. In any case, informed subjective judgment is absolutely essential.

The criterion for minimal mastery, expressed as some percent correct of the total number of test items, is explicitly under the examiner's control. In some testing situations, he may deem 70% correct as a minimal passing score, whereas when more critical skills are involved, he may want to observe at least 80% correct before calling an examinee a master. But as the model demonstrates, test length *interacts* with per correct required for "mastery" decisions. Specifically, as test length increases, classification accuracy increases, even when the *same* percent correct is maintained. In performance-based tests for example, where the cost of each item could be very high (such as field artillery or tank gunnery), the examiner is obliged to use the *minimum* number of trials, and so the minimal percent correct mastery criterion should be increased accordingly. Finally, the model has demonstrated that testing may be irrelevant in making mastery classification decisions if test length does not exceed some minimal number of items.

Appendix: A Computational Example for Three Mastery States

The following example illustrates the computations necessary for processing data with the Bayesian model. The values chosen for this example correspond to Figure 4. Assume that there are three states of mastery, and unequal prior probabilities for these three states. The educational decision-maker must provide estimates for the prior probabilities of master, $p(Mi)$. For this example let us assume the values to be: $p(M1) = .5$; $p(M2) = .3$; and $p(M3) = .2$. He must also provide estimates for the conditional probability of getting any given test item right, given each mastery state. The following values will be used as the conditional probability of getting an item right given a mastery state: $p(1|M1) = .8$; $p(1|M2) = .6$; $p(1|M3) = .5$. The conditional probabilities of getting an item wrong given a mastery state are: $p(0|M1) = .2$; $p(0|M2) = .4$; and $p(0|M3) = .5$.

First we need to calculate the probability that an item is answered correctly. For the overall population,

$$p(tj = \text{correct}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{correct}|Mi) = (.5)(.8) +$$

$$(.3)(.6) + (.2)(.5) = .68.$$

Likewise,

$$p(tj = \text{wrong}) = \sum_{i=1}^{S} p(Mi)p(tj = \text{wrong}|Mi) = (.5)(.2) +$$

$$(.3)(.4) + (.2)(.5) = .32.$$

We also need to obtain the set of conditional probabilities for the different mastery states given that an individual item was responded to either correctly or wrongly. The general equation is:

$$p(Mi|tj) = \frac{p(Mi)p(tj|Mi)}{p(tj)}$$

Substituting the above values yields: $p(M1|tj = \text{correct}) = (.5)(.8) \div .68 = .588$; $p(M2|tj = \text{correct}) = (.3)(.6) \div .68 = .265$; and $p(M3|tj = \text{correct}) = (.2)(.5) \div .68 = .147$. (Note that the sum equals 1.0.) Finally, $p(M1|tj = \text{wrong}) = (.5)(.2) \div .32 = .3125$; $p(M2|tj = \text{wrong}) = (.3)(.4) \div .32 = .375$ and $p(M3|tj = \text{wrong}) = (.2)(.5) \div .32 = .3125$. If 6 items were answered correctly on a 10 item criterion-referenced test, the following $\prod_{j=1}^{N} p(Mi|tj)$ values result:

$M1 = 3.9 \times 10^{-4}$;
$M2 = 6.8 \times 10^{-6}$;
$M3 = 9.6 \times 10^{-8}$.

Finally, the general Bayesian formula yields the conditional probability for each mastery state given the total test score. For example,

$$p(Mi|T) = \frac{(3.9 \times 10^{-4})}{(.5)^9 \left[\frac{(3.9 \times 10^{-4})}{(.5)^9} + \frac{(6.8 \times 10^{-6})}{(.3)^9} + \frac{(9.6 \times 10^{-8})}{(.2)^9}\right]} = .272.$$

Similar calculations yield $p(M2|T) = .473$ and $p(M3|T) = .254$.

References

Hershman, R.L. A rule for the integration of Bayesian opinions. Human Factors, 1971, 13, 255-259.

Novick, M.R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C.W. Harris, M.C. Alkin, & W.J. Popham (Eds.), Center for the Study of Evaluation Monograph Series in Evaluation III: Problems in criterion-referenced measurement. Los Angeles: U.C.L.A. Center for the Study of Evaluation, 1974.

FIG. 1.

P(M1) = .9 P(M2) = .1

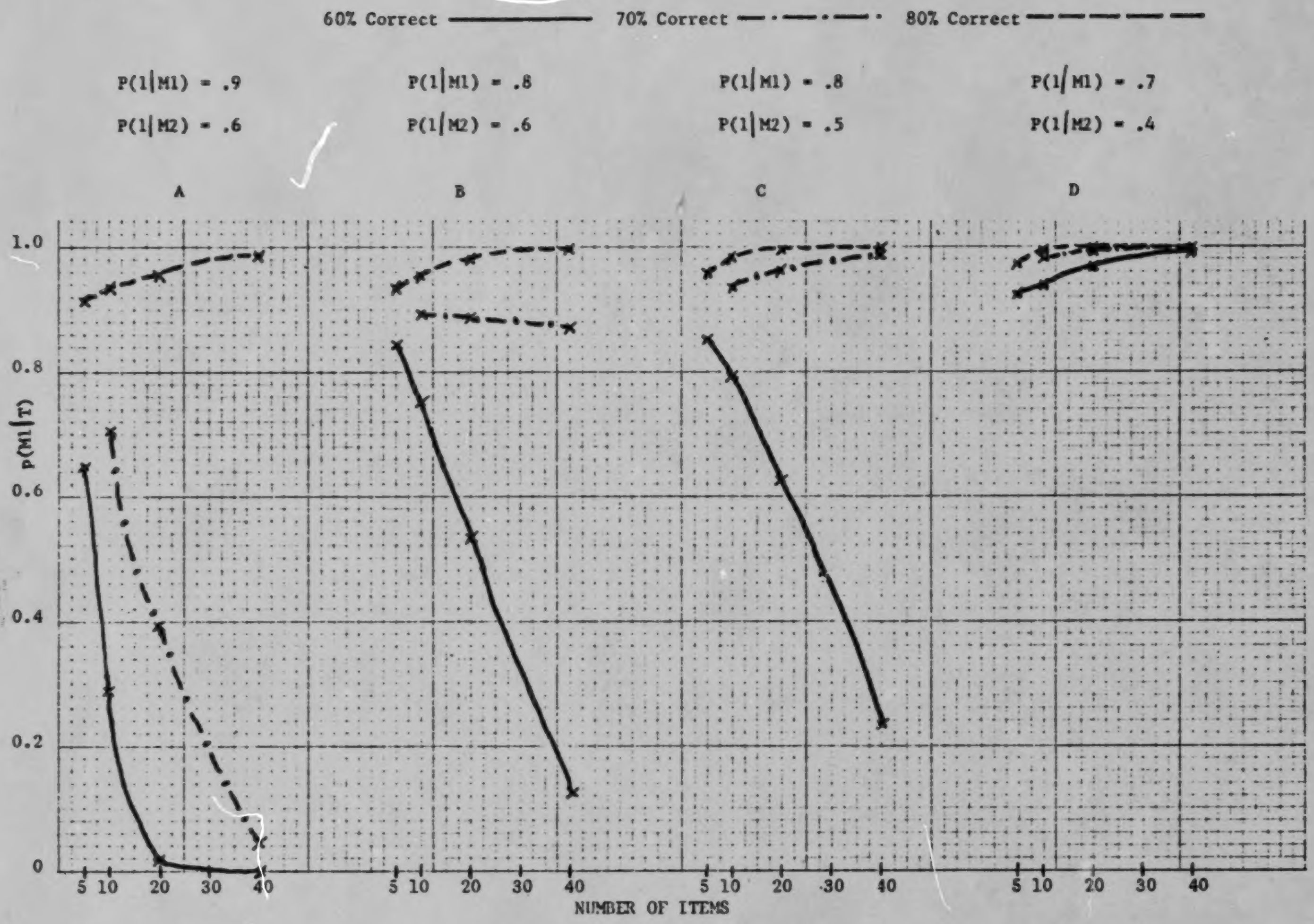


FIG. 2.

P(M1) = .7   P(M2) = .3

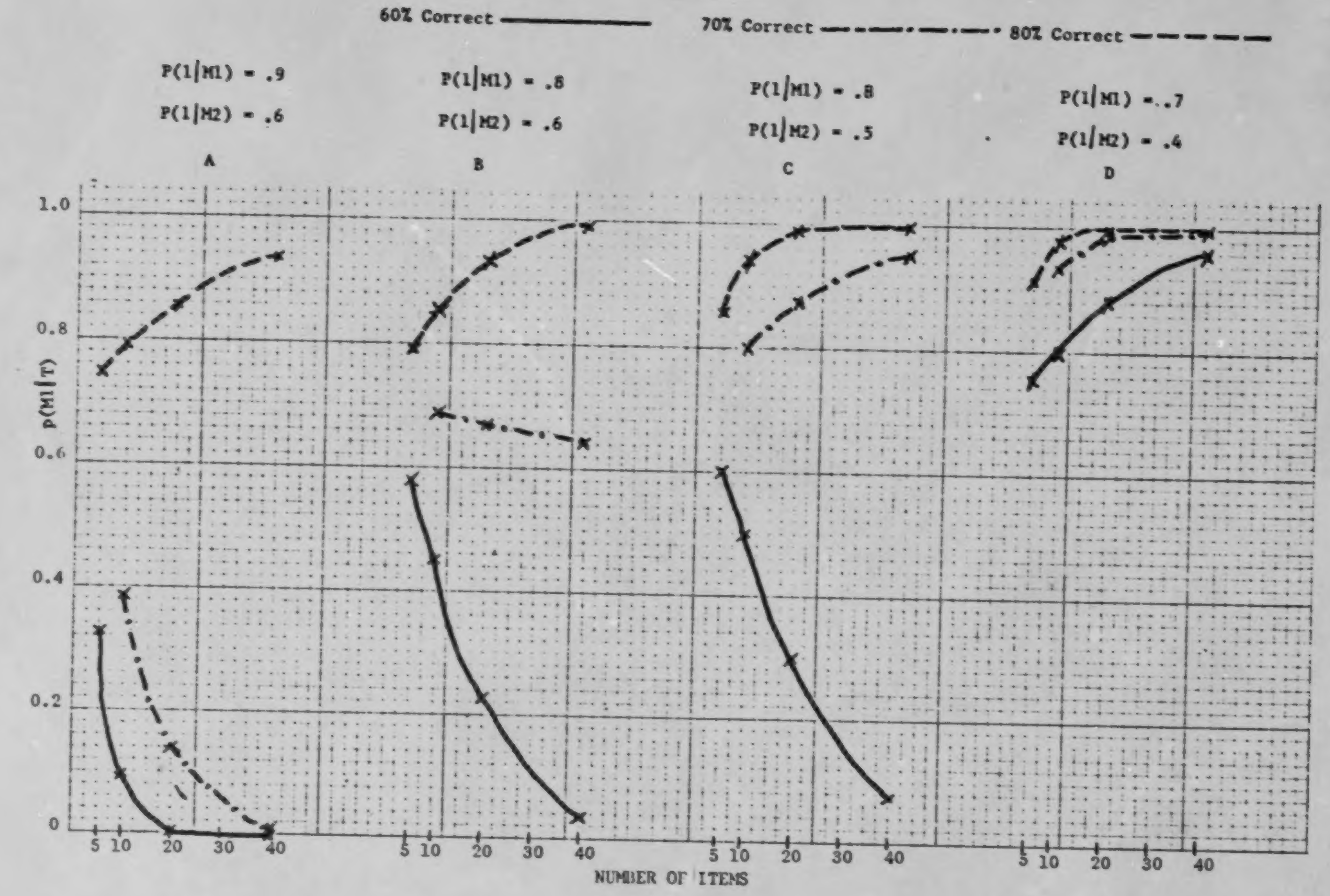


FIG. 3.

$P(M1) = .5$ $P(M2) = .5$

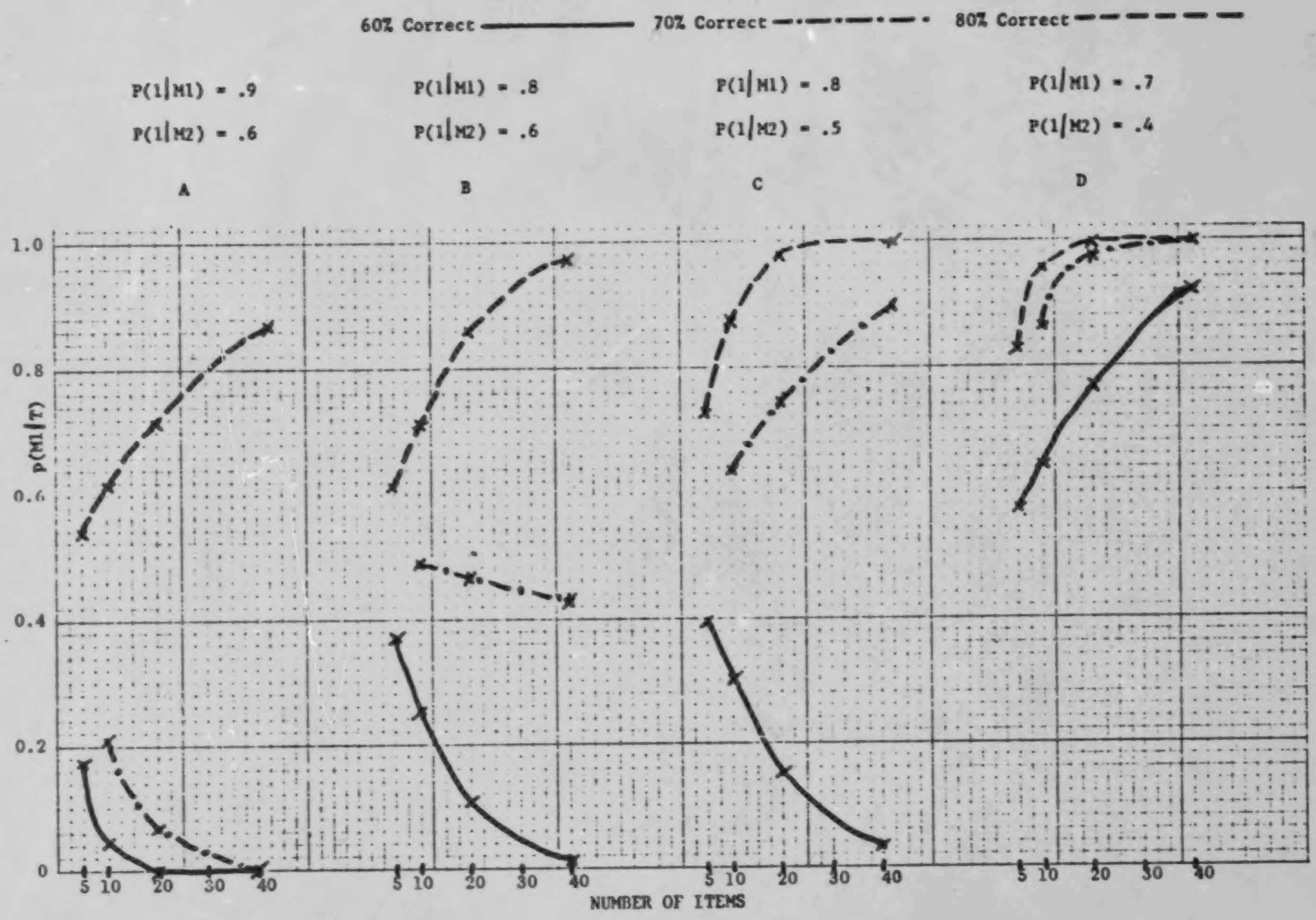


FIG. 4.

$\underline{P(M1) = .50}$. $\underline{P(M2) = .30}$ $\underline{P(M3) = .20}$

60% Correct ———————— 70% Correct —·—·—·—80% Correct — — — — —

$\underline{P(1|M1) = .8}$, $\underline{P(1|M2) = .6}$, $\underline{P(1|M3) = .5}$

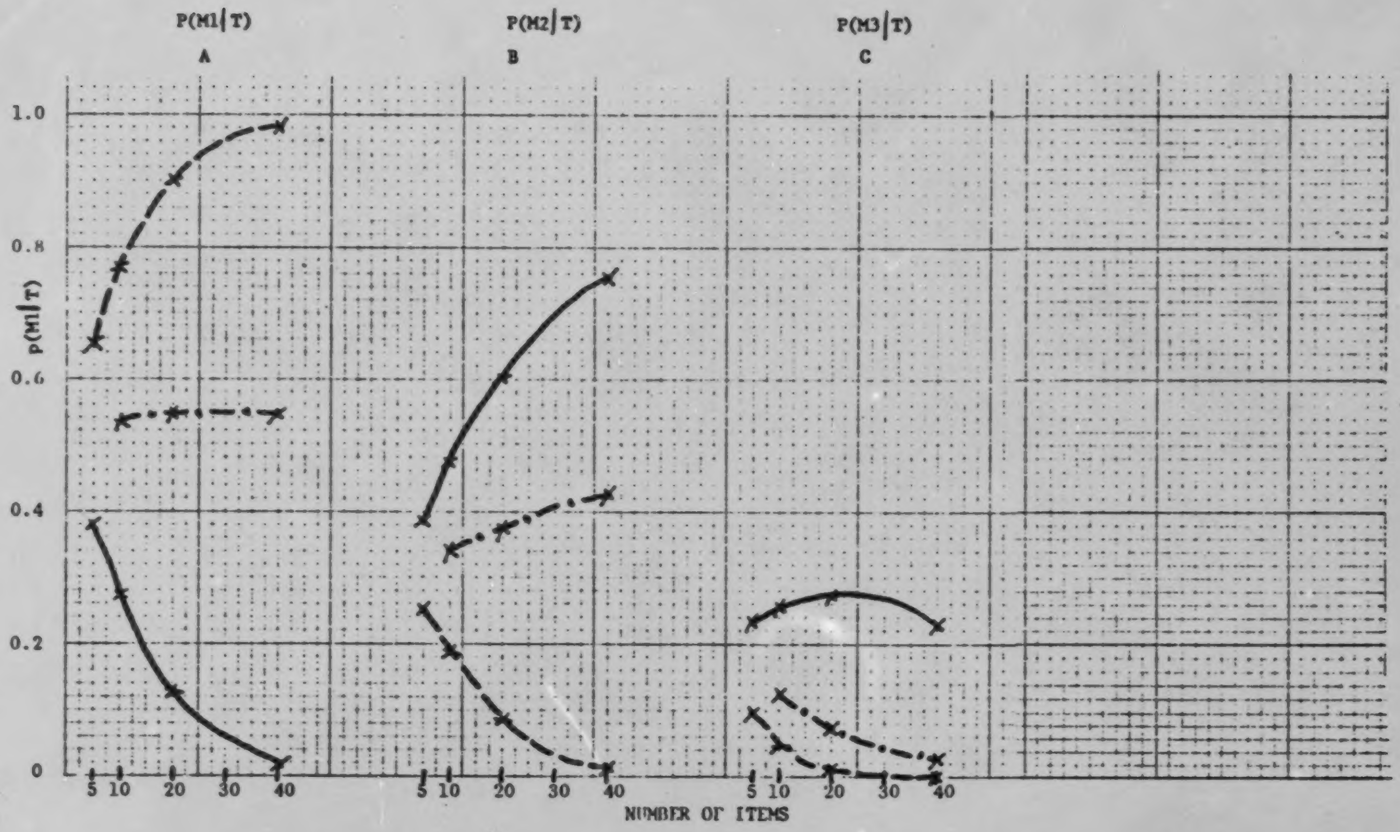


FIG. 5.

P(M1) = .50 P(M2) = .30 P(M3) = .20

60% Correct ———— 70% Correct —·—·—·— 80% Correct — — — —

P(1|M1) = .90, P(1|M2) = .80, P(1|M3) = .60

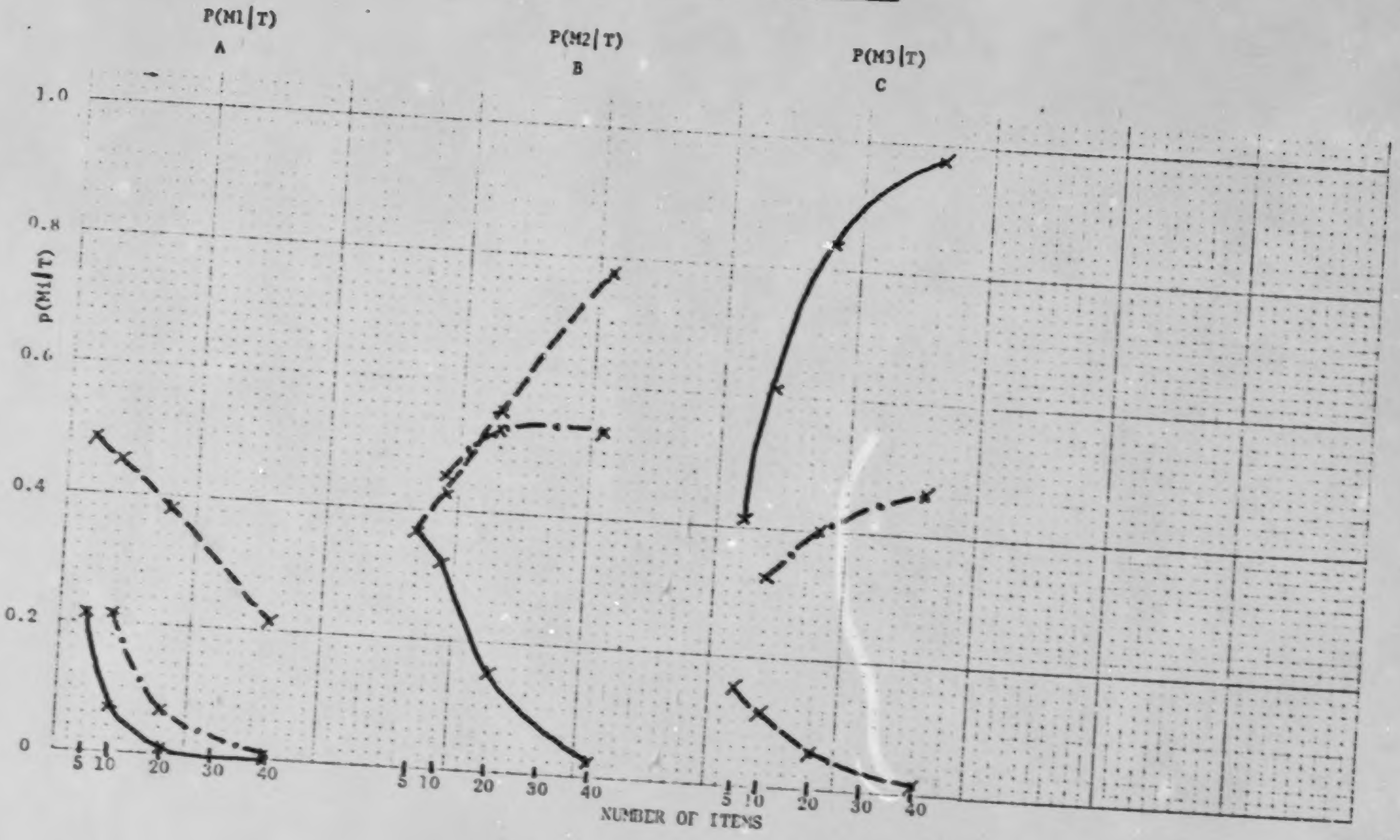


FIG. 6.

P(M1) = .25 P(M2) = .50 P(M3) = .25

60% Correct ———— 70% Correct —·—·— 80% Correct — — — —

P(1|M1) = .8, P(1|M2) = .6, P(1|M3) = .5

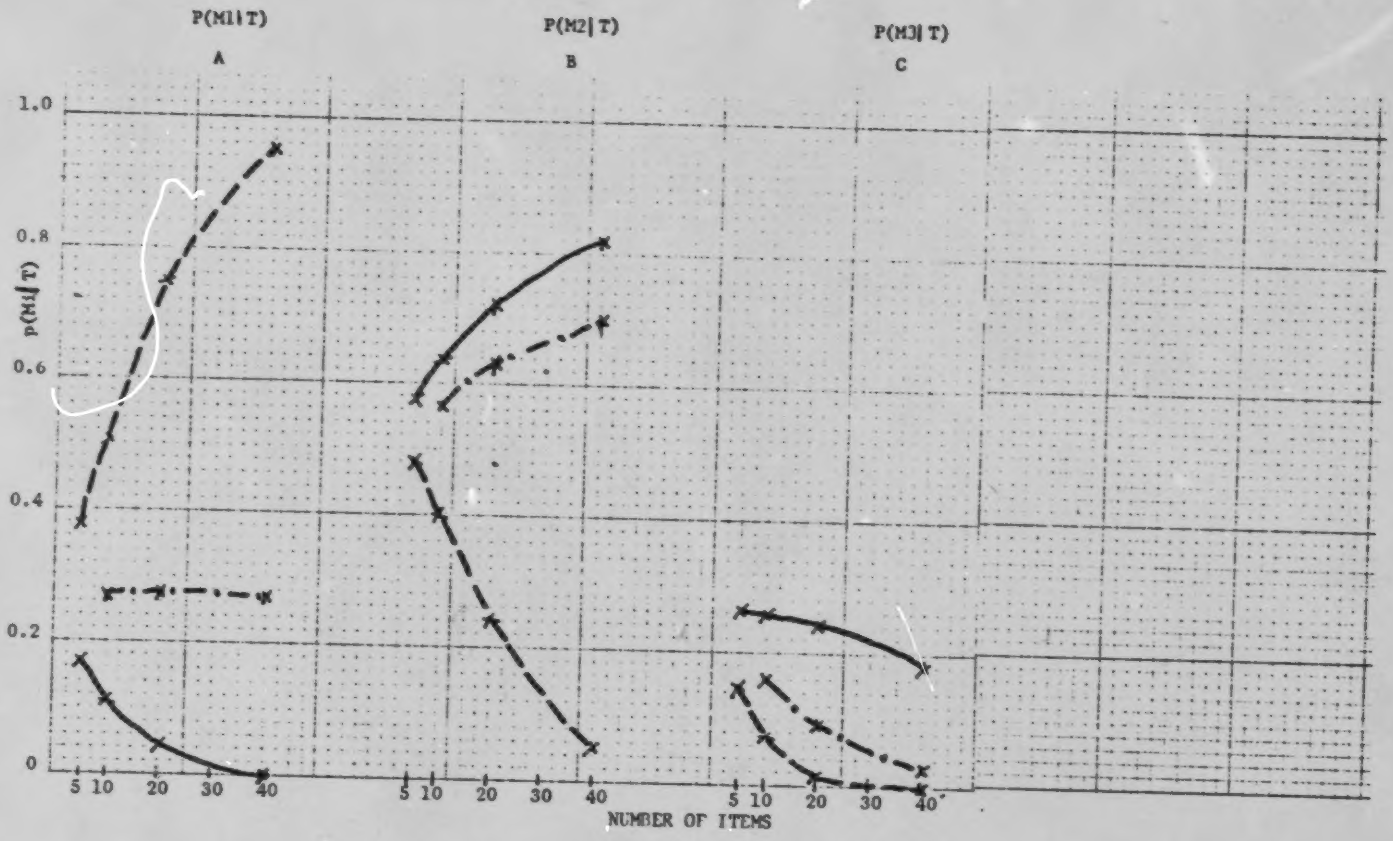


FIG. 7.

P(M1) = .25   P(M2) = .50   P(M3) = .25

60% Correct ———— 70% Correct —·—·—· 80% Correct — — — —

P(1|M1) = .90, P(1|M2) = .80, P(1|M3) = .60

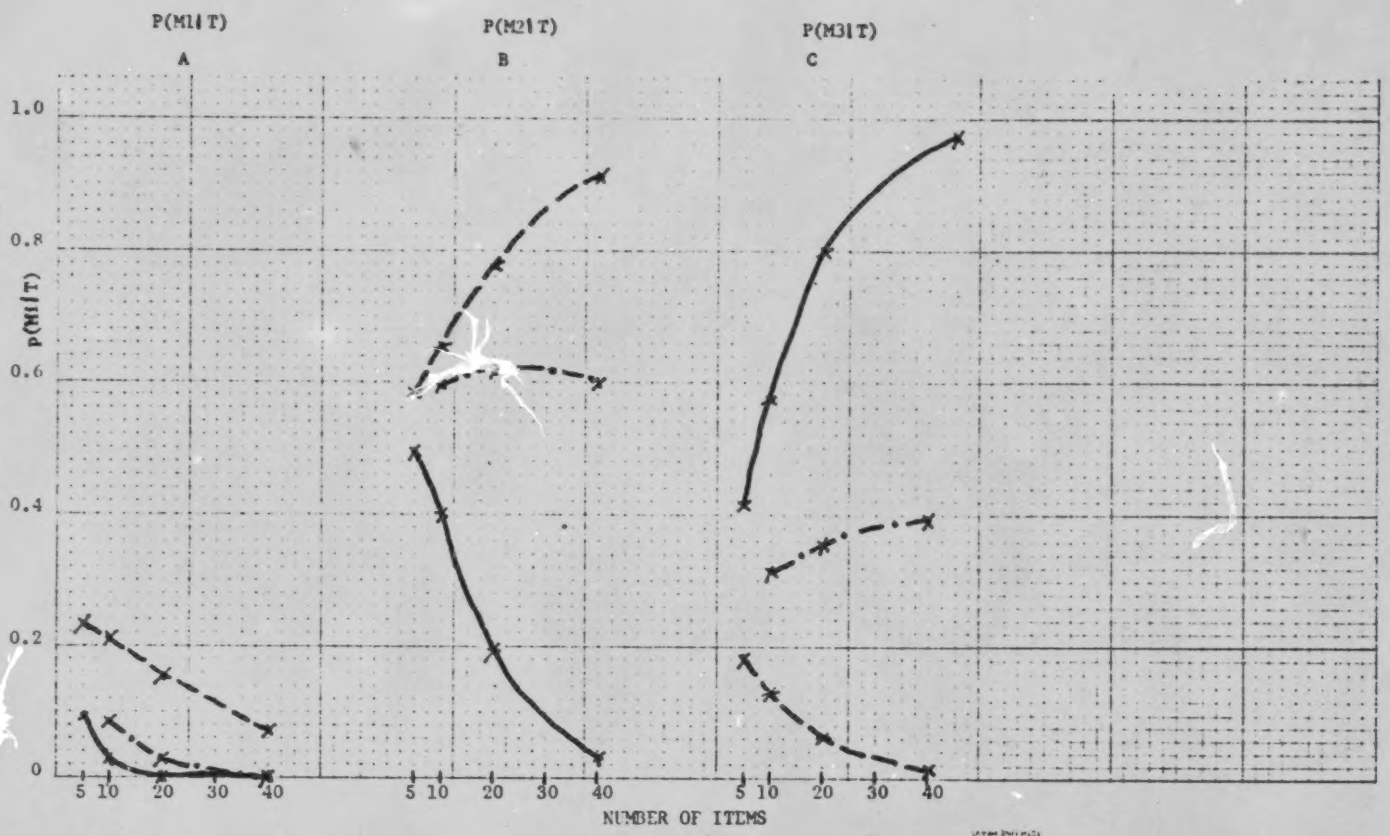


FIG. 8.

$\underline{P(M1) = .9}$ $\underline{P(M2) = .1}$

$P(1|M1) = .9,\ P(1|M2) = .6$

5 Item Test – – – – – 40 Item Test ——

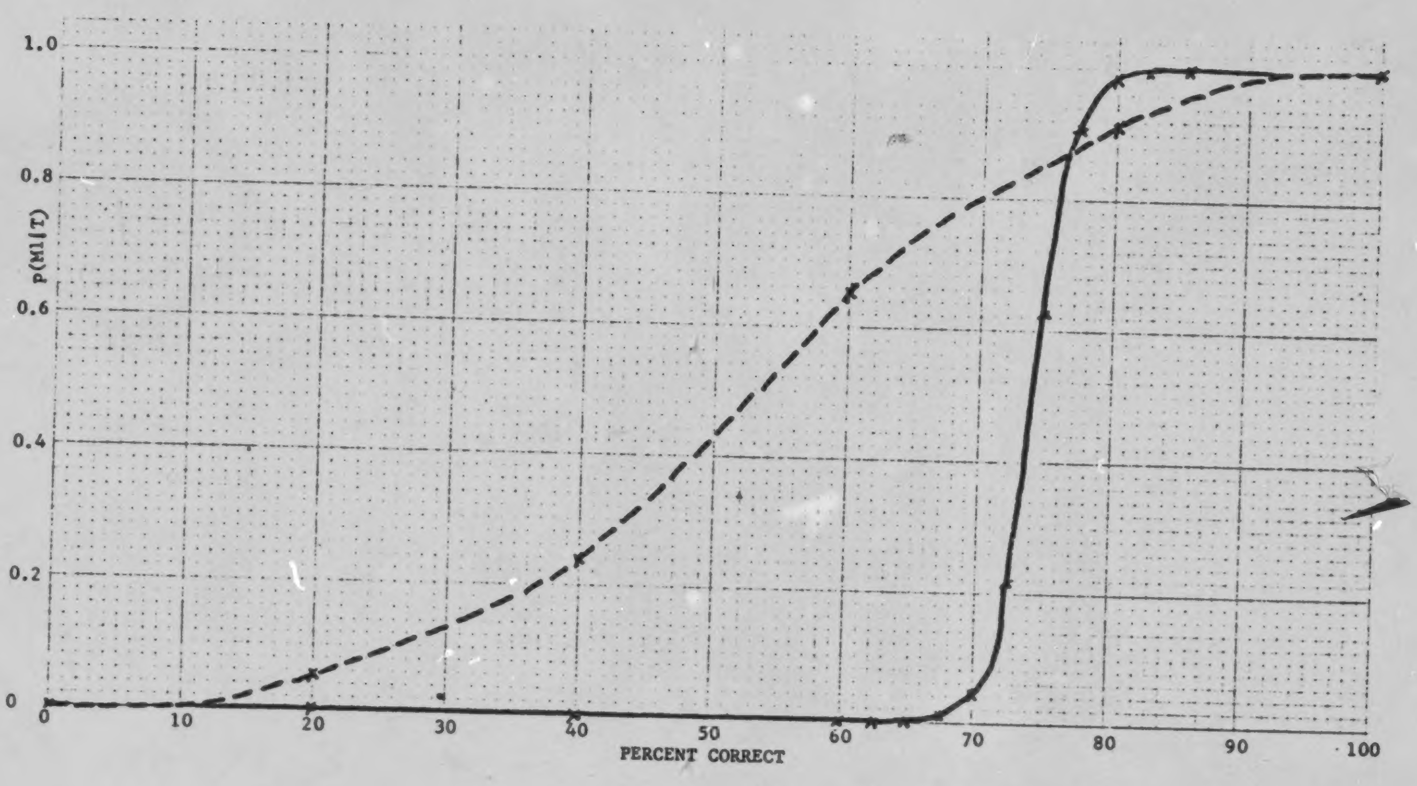
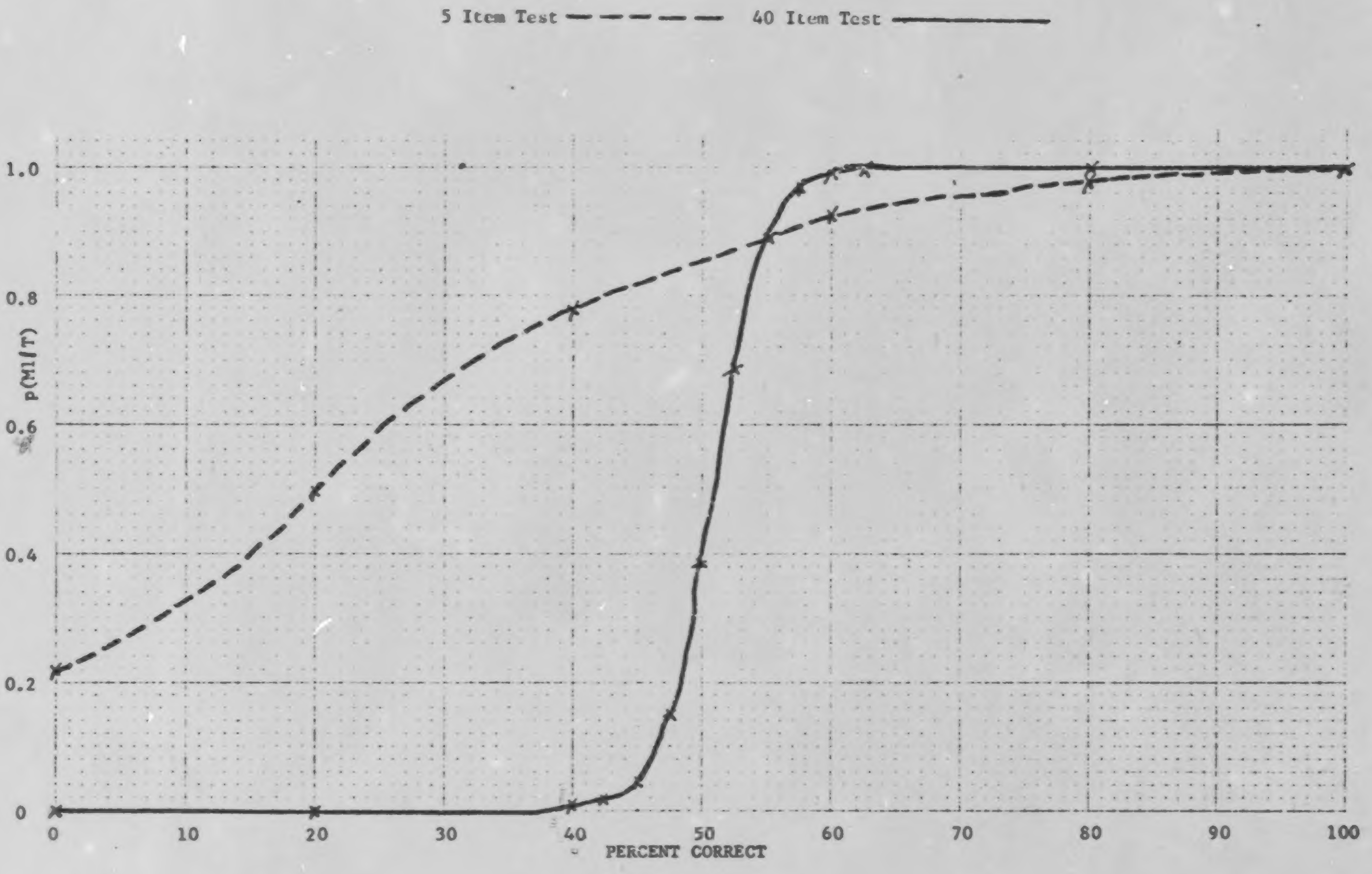


FIG. 9.

P(M1) = .9   P(M2) = .1

P(1|M1) = .7, P(1|M2) = .4



An Empirical Investigation of Four Criterion-Referenced Testing Models

Kenneth I. Epstein
Army Research Institute for the
Behavioral and Social Sciences

For the past two years, the Army Research Institute has been engaged in basic testing research under Project METTEST - Methodological Issues in the Construction of Criterion-Referenced Tests. Project METTEST was conceived to provide basic support for the Army's rapidly growing trend towards performance oriented training and testing. This paper summarizes part of the thinking which has evolved from METTEST.

One question that is present whenever test scores are considered is, What are the "true" scores that the fallible test scores represent? The problem is so complex that there is considerable discussion about what "true" score means (see, for example, Lord and Novick, 1968 pp. 39-44). Even if a definition is agreed upon, the estimated value of the true score given an observed test score will vary according to the particular measurement model used to evaluate the data.

The absolute value of true score has been relatively unimportant for classical measurement. This is because the primary purpose of classical testing is to rank order examinees consistently. For this purpose, the critical problem is not determining the true score, but rather it is determining the correlation between true scores and observed scores. The techniques of classical measurement are powerful because it is possible to determine this correlation without knowing the actual values of the true scores. It is possible to evaluate how well a given test yields scores that correlate highly with true scores and hence, how well the test will rank order examinees consistently.

Criterion-referenced testing does not allow for the luxury of placing the emphasis in test evaluation on correlation. In fact, it is possible for a good criterion-referenced test to correlate zero with true score. This might occur in a mastery learning context where all examinees have attained an equal level of ability and all examinees obtain the same test score. If one attempts to determine the correlation between a

The views expressed in this paper are those of the author and do not necessarily imply endorsement by the U. S. Army.

collection of equal observed scores and equal abilities, the value of the coefficient will be zero. The major purpose of criterion-referenced testing is not to rank order examinees consistently. Rather, it is to estimate the true capabilities of examinees to perform specific tasks. Hence, the problem of true score determination assumes critical importance.

This paper discusses four measurement models which have potential for evaluating the results of criterion-referenced tests. In particular, the paper compares the estimates each model yields for true scores given observed scores. The four models to be discussed are 1) a model based on observed proportion correct, 2) a binomial error model, 3) a Bayesian model, and 4) the Rasch one parameter logistic model.

The data for the analyses were collected as part of a study which evaluated tank gunnery training devices (Rose, et. al, 1975). The data consist of hit/miss scores recorded for 12 rounds fired from the main gun of the M60A1 tank at a moving target. 154 crews participated in the experiment. A summary of the data is given in Table 1. For the original experiment the 154 crews were broken into seven groups corresponding to different training programs. The reanalysis of the data for this paper ignored the difference in training since it is irrelevant to the present problem. At some future time it might be of interest to study whether test results are consistent across apriori defined different examinee populations.

The major requirement of a criterion-referenced test is that all items come from a well-defined domain (Millman, 1973). In other words, it is essential that all items in the test (or sub-test) measure the ability to perform the same objective. The twelve essentially identical rounds provide an ideal representation of a criterion-referenced test designed to evaluate gunner ability to hit moving 16' x 10' plywood tank silhouettes at ranges of approximately 700 and 750 meters with standard 105mm HEAT TPT ammunition. The twelve rounds represent a sample of an infinite number of similar rounds that could have been fired. Thus, it seems reasonable, and it is consistent with established measurement theory, to define the true score as the proportion of hits that would be obtained were the infinity of rounds fired. In other words, the twelve rounds represent a sample of trials chosen from an infinite population of possible trials. The task then is to infer from the performance on the twelve trial test what proportion of hits would be achieved if all trials were given. The following section of the paper discusses four procedures for accomplishing this task.

The first procedure uses the obtained proportion correct as the estimated true score. If we assume that the test is a random sample of trials from the domain of interest, and if the trials are dichotomously scored, and if we assume that the response to any given item is not dependent on the response to any other item (e.g., there is no learning occurring during testing, nor is there any fatigue or similar effect occurring), then it is appropriate to describe the data mathematically as a series of independent Bernoulli trials. In this case, the proportion of sample trials scored correct by an individual is an unbiased estimate of the proportion correct in the infinite domain for that individual. This model implies that all items are of equal difficulty for an individual. For example, if the proportion that would be correct in the whole domain is p for some individual, then the probability that he would respond to any randomly sampled trial correctly is also p. Notice that this does not imply that two individuals of differing ability will find the trials equally difficult. The more capable individual will find trials uniformly easy, the less capable individual will find the same trials uniformly difficult. Table 2, Column A shows the proportion correct corresponding to each obtained score on this 12 trial test.

The proportion correct can also be shown to be the maximum likelihood estimator of the true proportion correct (Lord and Novick, 1968, p. 88). The proportion correct has a mean value (over repeated sampling) of p, the true proportion correct and a variance of $p(1-p)/n$. Hence, particularly for small sample sizes, this estimate of the true score is probably not adequately reliable. Notice that the variance of the estimated proportion correct will not be a constant for all obtained scores. In fact, it is largest for the mid range of the distribution ($p = .500$ yields the largest variance. $.25/n$), and decreases to zero at the extremes. This is not an unreasonable result. We would expect very good or very poor examinees to respond in a highly predictable and consistent manner. It is far more difficult to make fine discriminations near the middle of the distribution. Hence, the practice in test construction to design items to be most sensitive to the range of abilities near the middle. Lord and Novick (1968) point out an interesting paradox associated with this type of analysis.

> "The standard error of measurement is smallest for examinees whose true scores are nearest one or zero. Should it not follow from this that the best measuring instrument is a test composed of items so easy that everyone will have a relative true score near one, or so hard that everyone will have a relative true score near zero?
>
> The answer to this question is that the effectiveness of a test as a measuring instrument usually does not depend merely on the standard error of measurement, but rather on the ratio of the standard error of measurement to the standard deviation of observed scores in the group. The more discriminating the test items, the larger will be the standard deviation of observed scores, other things being equal; and hence, the less will be the

> danger that true differences will be swamped by random errors of measurement and lost to view.
>
> The small standard errors of measurement that result when a test is made very easy are not beneficial because the standard deviation of observed scores for such tests is also small. This is most apparent in the limiting case when the test is so easy (or so difficult) that everyone gets a perfect (or a zero) score and both standard deviations are zero. Even though in this case there are no errors of measurement at all, such a test obviously is not discriminating among examinees and thus is not a useful measuring instrument (p. 252)."

One possible exception to the Lord and Novick statement might occur if a sample of examinees all of whom were complete masters (nonmasters) of the material took the test. Then one would have a test that did not distinguish among identical individuals, which is what is desired. However, the occasions when the examinee population is likely to be homogeneous enough for this to occur are probably infrequent.

A natural extension of the proportion correct model is the binomial error model. The binomial error model is more powerful than the simple proportion correct because the entire distribution of observed responses is included in the analysis. All of the assumptions discussed in relationship to the proportion correct model hold for the binomial error model. The major addition is the specification of the conditional distribution for observed score x for given true proportion correct T. This distribution is the binomial:

$$(1) \quad h(x|T) = \binom{n}{x} T^x (1 - T)^{n-x} \quad x = 0,1,\ldots n, \quad 0 \le T \le 1,$$

and n equals the total number of trials on the test.

Lord and Novick (1968) prove a very useful consequence of the model. "Under the binomial error model, if the observed score distribution is negative hypergeometric, then the regression of true score on observed score is linear (p. 517)." They then outline a procedure for estimating the parameters of the negative hypergeometric distribution from observed scores. The procedure was carried out for the tank gunnery data and the theoretical distribution was compared to the observed distribution. The value of χ^2 for this analysis was 3.451. Evaluation of this value with nine degrees of freedom yielded $.95 > p\ (\chi^2 = 3.451) > .90$. Since this represents adequate fit it is possible to proceed with the analysis assuming that the regression of true score on observed score is linear.

The regression function can be written

(2) $E(T|x) = \alpha_{21}\, x/n + (1 - \alpha_{21})\, \mu_x/n$, $x=0,1...n$, and μ_x = the mean of the observed scores, $\alpha_{21} = n/(n-1)\, [1 - \mu_x(n-\mu_x)/n\sigma_x^2]$ σ_x^2 = the variance of the observed scores. (Lord and Novick, 1968 p. 517, 521)

For these data the value of the regression function is

(3) $E(T|x) = .059\, x + .161$, $x = 0, 1, \ldots n$.

The estimated true scores, calculated using the above regression function are found in Table 2, Column B.

Comparing these results with those obtained under the proportion correct model shows that they are comparable, particularly in the mid range of the distribution. The differences are directly attributable to a regression effect in which the extreme values are regressed toward the mean of the distribution. What is happening, in effect, is that performance of the group is being used to help temper judgments about individuals. On the one hand, this point of view seems to contradict the philosophy underlying criterion-referenced measurement, that an individual should be judged on the basis of his ability and not compared with his peers. However, since in most cases the examinee population does have some characteristics in common and since extreme scores are suspect it makes sense to use all available data. This particular model is especially attractice because it has a built in validity check. If one is not successful in fitting the negative hypergeometric distribution to the data it implies that the regression is either not linear or that the binomial error model is not appropriate. In any case, it immediately warns the user to be careful of any interpretations he makes.

Lewis, Wang, and Novick (1973) have applied the same philosophy implied in the binomial error model, that all available information be utilized, to the development of a measurement model based on Bayesian statistical theory. The procedures are complex and require computerization for full utilization. The results of using their procedures to evaluate these data reveal some interesting and thought provoking implications of the Bayesian approach.

The procedure begins by mapping the observed scores into a new set of variables (g_j) using an arcsine transformation. The g_j are assumed to be normally distributed with mean $\gamma_j = \sin^{-1} \sqrt{T_j}$ and variance γ_j = $(4n + 2)^{-1}$, where γ_j is the transformed value of the true proportion of successes, T_j, and n in the number of test items. The assumption of normality is shown to be reasonable for tests of at least eight items. In addition to the observed data, the procedure requires that two additional parameters be specified by the user. These parameters describe the prior beliefs concerning the distribution of γ_j. The γ_j are assumed to be a random sample from a normal distribution with mean μ_Γ, variance ϕ_Γ The μ_Γ and ϕ_Γ are assumed to be independent, having a uniform and inverse chi-square distribution respectively. The first additional parameter that must be specified is the degrees of freedom for the inverse chi-square distribution. A recommended value for most practical purposes is eight. This value was used for the analysis described in this paper. The second additional parameter is also related to the inverse chi-square distribution. It is designated t and can be thought of as the length of a test that the user would consider to offer as much information as he now has (before testing) about the examinees. Thus, if very little is known, t will be small, and if the prior information is extensive, t will be large. Since relatively little prior information was available for the data described in this paper, the value of t chosen was three. Clearly, before this procedure can come into wide use, the relative importance of the t value to the final results must be investigated. Experience based on empirical applications of the procedure will also help in establishing practical guidelines for the use of the procedure. (For a more detailed discussion of the rationale behind this procedure see Novick, Lewis, and Jackson, 1973 and Lewis, Wang, and Novick, 1973).

Once the observed data have been transformed and the parameter values specified, the application of the model is relatively straightforward. Two alternative procedures are available. For test lengths up to 30 items, examinee groups up to 80 persons, and transformed score variances up to .05 (reasonable values for most applications), Wang (1973) has prepared tables of constants for carrying out the necessary calculations for their use. For larger sample sizes, Lewis, Wang, and Novick (1973) provide a procedure for carrying out the necessary calculations which involves solving a cubic equation. The later procedure was used for this example.

The most important result of the procedure is a regression equation for the posterior values of γ_j. For the example data the equation is,

(4) $E(\gamma_j \mid \phi_\Gamma, g) = .710\, g_j + .244$, where ϕ_Γ is the posterior variance

solved for by a cubic equation and the g_j are the transformed variables.

It is interesting to compare this equation with the regression equation obtained as a result of applying the binomial error model (Equation 3). The two equations have essentially the same form. While they cannot be directly compared since the Bayesian equation is written in terms of transformed variables one can compare the relative weights given the mean of observed data and the individual variables themselves. (The pure number in each equation reflects the weight applied to the mean). In the Bayesian approach the individual variables are weighted much more heavily than the mean. In the binomial error approach the mean seems to be relatively more important. This should result in the true proportion estimates found by the Bayesian approach to be regressed less toward the mean than the binomial error model results. These results are shown in Table 2, Column C. In general, a regression effect is seen for the Bayesian results but it is not as strong as the binomial error model results.

The heavy weight accorded the individual score values in the Bayesian approach is a direct result of the small amount of prior information. This makes intuitive sense, for if little is known about a group it seems unreasonable to put much emphasis in overall indices of group ability such as the mean. Had the prior information been more conclusive, more use would have been made of group data in determining the estimates of the true porportions correct.

The final model to be discussed is the Rasch one parameter logistic model. Superficially the Rasch model appears to be conceptually very different from the previously discussed models, however the differences are more apparent than real. The Rasch model hypothesizes that people are distributed on an underlying ability trait and further that their response to a test trial or item is governed purely by the ability of an individual and the difficulty of the trial. This is analogous to the probability of responding correctly to any given trial that underlies the response patterns of examinees in the proportion correct model, the true score distribution that is observed as a negative hypergeometric distribution in the binomial error model, and the posterior distribution of abilities in the Bayesian model. The Rasch model's strength lies in the fact that it is possible to calibrate a set of items with differing difficulties and administer different subsets of those items to different groups of examinees. The resulting estimates of individual abilities will be on the same scale and it will be possible to compare individuals regardless of the particular items chosen.

The data in this example were analyzed at the University of Chicago by Dr. Benjamin Wright. The results of the analysis showed that the twelve trials differed in apparent difficulty. The first trials seemed to be more difficult than the latter trials indicating that a gradual learning effect occurred during the test. Notice that this implies that the assumption that all items were identical is not strictly valid. It also calls into question the validity of interpreting the true score as the proportion of items that would be correct if all items were given. If items are not equally difficult is it not more important to know which items are responded to correctly? On the other hand, if this level of specificity is desired it will require far more extensive testing than seems practical. Two alternatives seem available. The first is to accept the approximation implied by the definition of true score as the proportion correct for all items. For most practical purposes this seems to be acceptable. The other alternative is to begin thinking in terms of latent traits. Such a transition will require that a large amount of empirical data be collected so that abilities expressed in terms of latent traits cna be given meaning in terms of observable behavior. For example, the data discussed in this paper were calibrated according to the Rasch model so that individuals scoring six hits were assigned an ability value equal to 0.00, and those scoring eleven hits an ability value equal to 2.45. (The entire set of Rasch ability values is shown in Table 2, Column E.) Stating abilities and minimum standards in terms of latent variables will allow for great flexibility in testing, but will require major efforts in interpretation.

The major purpose of this paper is to compare estimated true scores obtained by applying several measurement models. Therefore, it was necessary to sacrifice some of the information obtained from the Rasch analysis in order to obtain values on the same scale as the other values. This was accomplished by assigning the average difficulty of the twelve trials to each trial and applying the basic equation of the Rasch model.

The Rasch model states that the probability of a correct response to any given trial by a given individual is a function of the difficulty of that trial and the ability of the individual:

(5) $p = e^{(b-d)}/ 1 + e^{(b-d)}$, where b is the item's difficulty and d is

the person's ability. If the above equation is solved with b equal to the average item difficulty, the estimated average probability of a correct response is found. This value corresponds to the estimated true proportion correct found using the other models. These results are shown in Table 2, Column D. (Note that results are not shown for zero correct or twelve correct. This is because neither extreme score is used in the calibration and hence no ability estimates are obtained.) The estimates for the Rasch model are very similar to those found for the proportion correct model. This occurs because the Rasch model calibration attempts to find values for b and d which duplicate the observed data. The reliance on the group mean which is incorporated in the binomial error model and the Bayesian model is not found in the Rasch model.

Comparison of the estimated true proportion of hits for the four measurement models indicates relatively little difference among the models. For practical purposes, such as assigning individuals to mastery groups, there are not likely to be great differences when different approaches are used. Until more theoretical and empirical work has been completed it is not possible to make qualitative statements about the different approaches. However, some general strengths and weaknesses of the models can be identified.

The proportion correct model is clearly the easiest to apply. Calculations are minimal and the interpretation is straight forward. The approach has two weaknesses. First, none of the information about the group is incorporated. Group data is valuable in interpreting test results and should be considered. A second related weakness is that there is no direct connection between the estimated true proportions correct and the observed data. The calculations for the proportion correct model can be done without observed data. Thus, it seems to offer a good first approximation. If the observed data make sense when interpreted in terms of this model then it can probably be utilized. However, if examinees perform very differently than they would be expected to perform, then the model may be inappropriate.

The binomial error model shows, in effect, what happens when the proportion correct model is applied to observed data. Its strengths lie in its use of all the test information and the built in check on its fit to the data. An obvious weakness is that the model cannot be used if the negative hypergeometric distribution does not fit the observed data. In such cases the regression of true score on observed score is not linear. Techniques do exist for calculating the nonlinear regression but they require smoothing, estimation procedures, and tedious computations. Whether such measures are warranted is questionable.

The Bayesian approach will be most useful when significant prior information is available. For cases such as that presented in this paper the true power of the Bayesian approach will not be demonstrated. The Bayesian approach has the advantage of incorporating prior information into the analysis in addition to incorporating all the observed data. The major disadvantage of the Bayesian approach is that very often intuitive estimates of priors are in error. Such errors can lead to misleading results. However, the potential of the Bayesian approach for increasing the precision and efficiency of testing warrants that practical guidelines for its application be developed.

The Rasch model presents an opportunity for testing to pursue new directions. It has the potential for greatly increasing flexibility in testing. Like the binomial error model and the Bayesian model, the Rasch model relies on observed data in calculations, however once a calibrated item pool is available the person ability estimates are free of the particular item set used. This independence from the item set puts the major emphasis on the individual's ability. Thus, it seems philosophically more attuned to criterion-referenced testing. Weaknesses of the

Rasch model include possible problems in interpretation and the fact that, not all data sets will fit the model. If the data do not fit, it requires major revisions of the test or recourse to an alternative model.

The final choice of a model should be based on the needs of the testing program and the resources available to analyze the data. It is hoped that this paper will lead practitioners to more carefully consider their test results, whatever model they choose. In this way, perhaps the interpretations of test scores and the decisions based on them will be improved.

Table 1
Summary of Tank Gunnery Data

Observed Number of Hits	Frequency	Round Number	Proportion of Examinees Scoring a Hit
0	2	1	.429
1	2	2	.487
2	10	3	.526
3	12	4	.474
4	13	5	.500
5	16	6	.558
6	20	7	.545
7	16	8	.662
8	19	9	.578
9	14	10	.636
10	14	11	.604
11	11	12	.630
12	5		
	m=154		

Mean Number of Hits: 6.630
Variance: 8.457
Mean Proportion of Hits: .553
KR-20 = $k/k-1(1-\Sigma pq/\sigma_T^2) = .7136$

Note: The KR-20 reliability coefficient is included here not because the author necessarily advocates its use in evaluating criterion-referenced tests. It is included because it helps to describe the nature of this data and this group of examinees. A paper discussing some valid interpretations of classical measurement techniques for criterion-referenced tests is in preparation.

Table 2
Summary of Estimate True Proportion of Hits

Observed Score	A: Proportion Correct	B: Binomial Error	C: Bayesian Model	D: Rasch (Proportion)	E: Rasch (Ability)
0	.000	.161	.093	----	----
1	.083	.220	.201	.079	-2.45
2	.167	.279	.278	.161	-1.65
3	.250	.338	.342	.244	-1.13
4	.333	.397	.401	.330	-0.71
5	.417	.456	.469	.416	-0.34
6	.500	.515	.516	.500	0.00
7	.583	.574	.574	.584	0.34
8	.667	.633	.633	.670	0.71
9	.750	.692	.691	.756	1.13
10	.833	.751	.752	.839	1.65
11	.917	.810	.821	.921	2.45
12	1.000	.869	.921	----	----

References

Lewis, C., Wang, M.M., & Novick, M.R. Marginal distributions for the estimation of proportions in m groups. ACT Technical Bulletin No. 13. Iowa City, Iowa: The American College Testing Program, 1973.

Lord, F.M., & Novick, M.R. Statistical theories of mental test scores. Reading, Massachusetts: Addison-Wesley Publishing Company, 1968.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Novick, M.R., Lewis, C., & Jackson, P.H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Rose, A.M., Wheaton, G.R., Leonard, R.L., Fingerman, P.W. & Boycan, G.G. Evaluation of two tank gunnery trainers. Arlington, Virginia: U.S. Army Research Institute for the Behavioral and Social Sciences, in press.

Wang, M.M. Tables of constants for the posterior marginal estimates of proportions in m groups. ACT Technical Bulletin No. 14. Iowa City, Iowa: The American College Testing Program, 1973.

Wright, B. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

MATCHING JOB EDUCATION REQUIREMENTS WITH CANDIDATES' EDUCATIONAL ATTAINMENTS - A PILOT METHODOLOGICAL STUDY

By

Squadron Leaders William J. Watson & Kenneth Goody (RAAF/USAF)
Occupational & Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland AFB, Texas 78236

I. INTRODUCTION

As with other military organizations, the United States Air Force promulgates education qualifications for assignment to and service in the various officer utilization fields (AFM 36-1 and AFM 50-5). Some of these qualifications are listed as being desirable, others as mandatory. The implication is that some educational backgrounds are more suitable for service in particular utilization fields than are others, an implication with face validity. This paper identifies limitations of the present system for specifying these qualifications, and proposes an alternative methodology.

II. ANALYSIS OF THE PROBLEM

Typically, educational requirements for officer utilization fields are expressed in terms of possession of a particular degree, sometimes also specifying particular subjects that should be included in that degree. To illustrate, AFM 36-1 gives the education qualification for the pilot and navigator specialties as:

> "Bachelor of Science degree with appropriate courses in physical science, mathematics, administration, and management is desirable."

To illustrate some of the limitations of the present system, Table 1 was prepared from data gathered in this study. It breaks down five actual college degrees by the number of semester hours in the various areas of study relevant to the stated desirable education qualifications for pilots and navigators. It includes the type of degree awarded, and the grade point average (GPA) when available.

Table 1. Content of Five Actual College Degrees

Degree Identification Number	1	2	3	4	5
Type of Degree	BS	BS	BS	BA	BA
Semester hours in Mathematics	28	15	9	6	39
Semester hours in Engineering	51				
Semester hours in Physics	11				8
Semester hours in other physical sciences	8	15	3	6	
Semester hours specifically designated Administration or Management	3	18	6	4	3
Semester hours in Business Studies	19	14			20
Semester hours in Military Studies	15	30	38	41	15
Semester hours in other subjects	18	44	83	72	50
Grade Point Average (GPA)	3.22	2.06	2.12		3.25

The first three degrees in Table 1 fit the stated education qualifications for pilot/navigator; the last two have the desired subjects but they are arts degrees, not science degrees. The third and fourth degrees are virtually identical, but one is an arts degree and the other is a science degree. This illustrates the first limitation of the present system. What is a science degree from one college may be an arts degree from another. This being the case, the validity of specifying a science degree in preference to an arts degree must be questioned.

A second limitation of the present system is that it permits only dichotomous decisions. Either a candidate has the stated qualification, or he does not. There is no consideration of degree of suitability. For example, the first three degrees in Table 1 all meet the requirements for service as a pilot or navigator, and so each is presumably equally suitable. The low Mathematics/Physics content of the third degree, and the barely passing GPAs of the second and third, apparently have no relevance.

The problem of using college education for officer assignment is further compounded by the fact that the only documentary evidence of educational attainment is the candidate's college transcript. This can be a highly confusing source of information, since the several thousand colleges in the United States have no standard terminology in the transcripts. They use the same names for different courses, different names for the same courses, different standards and scales for assessing and reporting quality of academic attainment, and so on. The transcript, while it is a suitable vehicle for comparing the academic attainments of a few candidates, is totally unsuited for use on the scale needed for Air Force officer assignment.

The officer utilization field assignment process might be viewed as one of matching "pegs" to "holes". There is a pool of candidates ("pegs") comprising officers entering the service, and serving officers being considered for re-assignment. Each member of this pool has a unique combination of attributes, one of which is educational attainment. On the other hand, there are vacancies in the various utilization fields ("holes") that are to be filled. Each of these has its own unique combination of characteristics which imposes demands on the mandatory and desirable attributes of potential assignees, including educational background. A one-to-one match between the education qualifications held by the various members of the pool of candidates and the educational requirements of the various "holes" is impossible. In fact, an unpublished study by Judy (1971) revealed that for many utilization fields this match is quite poor. Many candidates are being assigned to fields for which they do not meet the educational requirements stated as being desirable.

To avoid assigning officers to fields where their college education has little practical value and for which they have little academic affinity, and to make the best match of existing talent to established vacancies, there is a need for a systematic methodology for expressing the educational background of a candidate in terms of degree of education suitability for each of the various utilization fields. The coarse, ambiguous, dichotomous system presently used does not meet this need. The research reported in this paper establishes the feasibility of an alternative methodology that might. Its specific aims were:

(a) To design an "education profile" by means of which both the quantity and quality of an officer's college education, as evidenced by his college transcript, could be condensed into a standardized, manageable display.

(b) To measure the reliability with which college transcripts could be translated into the profile format.

(c) To develop a set of profiles from the transcripts of a sample of officers recently commissioned into a wide range of officer utilization fields.

(d) To have this set of profiles rated by a panel of "experts", on the educational suitability of each officer for assignment to the Administrative Officer Utilization Field (AFSC).

(e) To estimate the level of agreement between panel members on the relative educational suitability of these officers.

(f) To "capture" the policy of the panel as a mathematical model, thereby providing a system capable of automatically computing an "education suitability index" for candidates for the utilization field.

III. THE PROPOSED SYSTEM

The Administrative Officer Utilization Field was chosen as the vehicle for developing and testing the methodology. There were two reasons for this decision. The first was administrative convenience - there are a large number of administrative officers assigned at Lackland AFB, Texas, where the Air Force Human Resources Laboratory, Occupational and Manpower Research Division is located. Secondly, it was hypothesized that if a reasonable level of interrater agreement existed between the judges in such a heterogeneous field, there was a good chance that judges in more homogeneous areas would also show agreement in assessing education suitability for their field.

Designing an "Education Profile"

Since the typical college transcript was unsuitable for use by judges in assessing the relative value of the college education of large numbers of candidates, a simplified, condensed version, termed an "Education Profile" was designed. The experimental profile was designed to the following specifications.

1. It should display all essential information of the candidate's college education in a simple standard format.

2. Data should be quantified whenever possible.

3. It should permit fine discrimination in areas of Air Force interest.

4. It should be expandable to form the basis of a permanent education record for the duration of an officer's career.

The developed profile form is shown in Appendix A, together with the definition of each of the 20 "education headings" used, and sufficient examples and exclusions to permit practical use. These 20 education headings represent the smallest set by which all college courses of interest to the Air Force could be described. The specificity of the headings varies depending on Air Force requirements in that area, e.g., the heading "Aerospace Engineering" is much more specific than the heading "Arts, Fine Arts and Humanities" heading. This is because qualifications in aerospace engineering are much rarer than general arts-type qualifications, and as they relate to an established Air Force need, there is a requirement to be able to identify them more precisely.

Appendix B shows the Profile Score Sheet, used for the clerical process of counting, for entry on the Education Profile, the number of semester hours each transcript revealed under each education heading. In assigning courses to education headings, the first heading, Calculus, was considered first; all courses on the transcript were examined, and those falling within the heading Calculus (as defined) were scored; from the remaining courses, those falling within the definition of the second heading, Probability/Statistics were scored; all other mathematics courses were then designated "Math Other" under the third heading. In like manner, each of the 20 headings was applied in sequence to each remaining course on the transcript, until finally all courses were assigned to an education heading.

By considering each transcript course title in this manner, it was possible for different judges to consistently assign even vaguely named courses to the same education heading. Three behavioral scientists from the Occupational and Manpower Research Division, including one of the authors, each scored 50 transcripts using the Profile Score Sheet, and the instructions for its use shown in Appendix C. Applying the intraclass correlation technique (Lindquist, 1953) for the three raters for each of the 20 headings gave an average interrater agreement of .95. This interrater agreement generally indicates that there should be no major problem in converting information from the transcript to the profile format in an operational system.

Education Profiles were then prepared from the average scores on the Profile Scoring Sheets for these 50 transcripts by adding Degree, Major, College and Grade Point Average (GPA), and re-arranging the education headings into the five academic areas Mathematics, Engineering and Physics, Other Sciences, Officer Managerial Studies, and General Studies. A typical completed profile is shown in Appendix D. A further 50 profiles were then prepared from other transcripts, giving a total of 100 in all. Some profile data were unavailable as some of the college transcripts did not specify any academic major or Grade Point Average (GPA). GPA was included only for those colleges using a 4-point system. This was the commonest, but far from the only system encountered; and 30% of the final profiles lacked GPA information. The scores on all profiles were converted to semester hours using the conversion factors one quarter hour = .75 semester hours, 1 unit = 4 semester hours. These factors applied for most colleges using quarters or unit systems, but in a few cases the precise factor was uncertain.

Education Suitability for Assignment as an Administrative Officer

Twenty three officers of grade second lieutenant through major currently serving in the 70XX Administration Utilization Field at Lackland Air Force Base, Texas volunteered to assess the sample of 100 profiles and rate the candidates on education suitability for assignment as an administrative officer. These officers comprised the "panel of experts" for this pilot study. Each was provided with the 100 profiles (randomly sorted to avoid context effect), a letter briefly outlining the project, a copy of the education heading definitions and a rating booklet containing instructions on what was required.

Each rater was asked to assign each profile to one of nine groups based on the education suitability it displayed. The scale for defining the groups ranged from Group 1 - most unsuitable - education has very little value (for assignment as administrative officer) to Group 9- most suitable - difficult to improve, with all points between defined. All profiles in Group 1 then automatically were rated 1, Group 2 rated 2 and so on. It was felt that this technique of sorting as part of the rating process would give the raters more opportunity to revise their ratings and permit finer discrimination than a simple "rate each profile in turn" system. Most raters reported taking about 1 1/2 - 2 hours to complete the rating task, although a few reported times as high as 6 - 12 hours.

The first stage in the analysis was to determine the agreement among panel members concerning the relative suitability of the candidates' education. Analysis revealed that nineteen of the 23 raters correlated positively with the mean rating. Two raters showing low negative correlation were deleted from the sample, while two raters with high negative correlation were ratained after adjusting the scale reversal involved. The raw scores were standardized by adjusting to a mean of five and a standard deviation of one, and the inter-rater agreement was measured using the intraclass correlation technique (Lindquist, 1953). The results were:

$$n = 21 \qquad R_{11} = .427 \qquad R_{kk} = .939$$

Capturing Rater Policy

This interrater reliability shows that with the profile information provided, administrative officers exhibit a high level of agreement on the relative educational suitability of anonymous graduates for assignment to the administrative utilization field. The next stage in the analysis was to determine what factors had been used by the panel members in reaching their decision.

A set of 50 variables was defined to quantify the data on the profiles. These variables, listed in Appendix E, include the college major, profile scores in semester hours, grade point average, and 13 "college quality" scores (Astin, 1965), five of which relate to the academic quality of the freshmen entering the particular college and eight to the college environment. The Astin variables were not explicitly available to the raters as a display on the profile, but were implicitly available depending on the rater's knowledge of, and attitude toward, the college named on the profile.

The mean adjusted rating on each of these 100 profiles was accepted as the best measure of suitability for service in the administrative career field (criterion), and the 50 variables quantifying the data in the education profiles are potential predictors of this criterion. However, to use all 50 of them as predictors in a multiple regression equation would be futile. Regression weights so obtained would be impossible to interpret, particularly as the linear dependencies between the predictors prevent a unique solution. Futhermore, the use of 50 predictors to compute a best fit regression equation from 100 cases would be "over-fitting", capitalizing too much on chance relationships.

The approach taken was to seek much smaller groups of linearly independent predictors that could efficiently predict the criterion. The resulting mathematical model would be relatively easy to interpret, and would establish the feasibility of a practical methodology for evaluating suitability of educational qualifications for service in particular areas of specialization. One such set, containing 10 predictors accounted for over 93% of the criterion variance. Details of the regression equation are presented in Table 2.

Inspection of the standard weights of the various predictors in the regression equation reveals which aspects of the education profile make positive contributions to the criterion, and which make negative ones. A large number of semester hours in Officer Managerial Studies Area are most beneficial, particularly in Business Studies (standard weight = .58). In this group Law, with a standard weight of .09, makes the least contribution. The various elements of this group do make individual contributions; pooling them together as a single variable reduces the squared multiple regression correlation coefficient from .93 to .89. Numerous hours of Engineering and Physics (standard weight = -.26) are a distinct disadvantage. Except for Social, Behavioral, Educational or Political Sciences which are desirable (standard weight = .34), "Other Sciences" also have a negative contribution (standard weight = -.19). However, the effect of these two areas is not as great as the standard weights would indicate because of the relationship between the two variables. The effect of the Grade Point Average (GPA) is another example of compensating weights. The "Existence of a Grade Point Average" predictor must be included to avoid undue penalty to the 30% of the candidates without a Grade Point Average.

IV. CONCLUSIONS AND RECOMMENDATIONS

This research effort has established that college transcripts can be condensed and quantified into education profiles with a high degree of reliability. Experts can use such profiles to rate candidates on educational suitability for service in the Administrative Officer Utilization Career Field. These ratings can be duplicated with a high degree of accuracy using only a few quantified aspects of the education profile, the best predictors being number of semester hours completed in certain fields of study plus the candidate's GPA if he has one. This research therefore establishes the feasibility of a methodology that translates college transcripts into education suitability indices, at least for some utilization fields.

This was a pilot methodological study. The restrictions created by using a non-random sample of raters is acceptable in this context, but would not be acceptable in developing an operational instrument. This research has studied one utilization field only; further research is necessary to confirm that it could be applied to any utilization field. To have the same 100 profiles rated on suitability for pilot and navigator utilization fields and repeating the analysis would be particularly fruitful follow-on research.

The operational instrument that could be developed from this research would condense college transcripts into the standard education profiles, and then compute an educational suitability index for various officer utilization fields. This index could then be used, along with other pertinent data, for assignment to utilization fields.

Table 2. Regression Equation Using 10 Education Variables to Predict Suitability for Service in the Administrative Career Field

Variable	Variable Description	Mean	Standard Deviation	Standard Weight	Regression Weight	Predictor/ Criterion Correlation
21	Social, Behavioral, Educational or Political Sciences	29.63	22.05	.340	.0155	.171
23	Basic Communication Skills	7.83	5.32	.259	.049	.322
24	Administration or Management	1.56	2.92	.238	.082	.590
25	Business Studies	8.40	13.52	.576	.043	.742
26	Law	1.09	3.66	.091	.025	.407
27	Military Studies	13.41	7.20	.171	.024	.058
31	Engineering & Physics Area Score (Sum of Variables 13-18)	16.13	23.43	-.262	-.011	-.507
32	Other Sciences Area Score (Sum of Variables 19-22)	48.67	25.79	-.190	-.007	-.085
36	Existence of Grade Point Average (GPA) (0 or 1)	0.70	0.46	-.150	-.327	.235
37	Grade Point Average (GPA)	1.99	1.36	.299	.222	.206

Regression Constant = 3.651

Multiple Correlation Coefficient (R) = 0.9651

Multiple Correlation Coefficient Squared (R^2) = 0.9315

REFERENCES

Air Force Manual 36-1. *Officer Classification Manual.* Washington: Department of the Air Force, May 1969.

Air Force Manual 50-5. *USAF Formal Schools Catalog.* Washington: Department of the Air Force, September 1974.

Astin, A.W. *Who goes where to college.* Chicago: Science Research Association, 1965.

July, C.J. The need for better identification of relevant elements in the educational background of Air Force officers. (Unpublished paper, 1971)

Lindquist, E.F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin Co. 1953, 359-361.

CASE NUMBER: ____________

APPENDIX A

OFFICER EDUCATION PROFILE

	DEGREE	MAJOR	COLLEGE
1			
2			
3			

EDUCATION PROFILE		(Semester Hours)	
MATH	Calculus		
	Probability or Statistics		
	Other Mathematics		
Engineering and Physics	Aerospace Engineering		
	Mechanical Engineering		
	Electrical or Electronic Engineering		
	Civil or Architectural Engineering		
	Other Engineering		
	Physics		
Other Sciences	Other Physical Sciences		
	Biological, Agricultural or Medical		
	Social, Behavioral, Educational or Political		
	Computer Programming or Use		
Officer Managerial Studies	Basic Communication Skills		
	Administration or Management		
	Business Studies		
	Law		
	Military Studies		
General Studies	Arts, Fine Arts, Humanities		
	Miscellaneous		

Grade Point Average		TOTAL	

EDUCATIONAL CATEGORIES

DEFINITIONS

No.	TITLE	DEFINITIONS/EXAMPLES/EXCLUSIONS
1.	Calculus	Courses specifically designated Calculus. Includes composite titles such as Calculus 3 with Vectors, Analytical Geometry/Calculus I.
2.	Probability or Statistics	Courses where these topics are specifically designated or strongly implied. Includes Introductory Statistics, Business Statistics, Measurement in Education, Psychometrics, etc.
3.	Other Mathematics	Courses in Pure or Applied Math other than those previously listed. Includes various algebras and geometries, complex variable, mechanics, statics, dynamics, etc., also courses listed as Mathematics without any clarification, e.g., Math III. Excludes applied mechanics or dynamics, e.g., Thermodynamics (Physics); fluid mechanics, mechanics of vibrations (Mechanical Engineering).
4.	Aerospace Engineering	Courses studying aeronautical, astronautical or aerospace vehicles or systems. Includes aerodynamics.
5.	Mechanical Engineering	Courses specifically designated, or courses in design and construction of non-flying machines. Includes fluid dynamics, vibration mechanics, etc.
6.	Electrical or Electronic Engineering	Includes computer design
7.	Civil or Architectural Engineering	Courses specifically designated, or courses in design and construction of buildings, towns, etc.
8.	Other Engineering	Includes all engineering courses not previously listed, e.g., Chemical, Industrial, Human Factors, Systems, Safety, Drafting, Engineering I (unspecified), Engineering Laboratory (unspecified), etc.

Appendix A (Continued)

No.	TITLE	DEFINITIONS/EXAMPLES/EXCLUSIONS
9.	Physics	Includes Engineering Science, Astronomy, Meteorology, Thermodynamics, etc.
10.	Other Physical Sciences	Includes all physical sciences other than Physics, e.g., Chemistry, Earth Science, Geology, Photography, etc.
11.	Biological, Agricultural Medical Sciences	Includes Pharmacy
12.	Social, Behavioral, Educational, or Political Sciences	Includes Anthropology, History, Geography, Political Science, Sociology, Psychology, Education, etc.
13.	Computer Programming or Use	Courses specifically designated. Includes Elements of Computer Programming, Mechanical Languages, etc. Excludes Computer design.
14.	Basic Communications Skills	Courses in written or oral communication skills relevant to AF jobs. Includes English composition, report writing, journalism, Freshman English (unless the transcript shows the emphasis to lie in the Fine Arts area), Fundamentals of Speech, etc. Excludes Fine Arts studies such as literature, poetry, drama, etc.
15.	Administration or Management	Courses specifically designated Administration or Management in any discipline except Military Studies. Includes Personnel Management, Engineering Management, Public Administration, etc. Excludes AFROTC Management or Administration courses.
16.	Business Studies	Includes Accounting, Economics, Marketing, Advertising, etc.
17.	Law	Excludes Military Law
18.	Military Studies	AFROTC, etc. Includes Military Administration, Management, Law.
19.	Arts, Fine Arts, Humanities	Includes Literary Studies, Foreign Languages, art, music, philosophy, religion, etc.
20.	Miscellaneous	Courses which cannot be rationalized within the above categories. Includes Golf, Commercial Flight I, Physical Conditioning, etc.

APPENDIX B

Scorer No. ___________ Transcript No. __________

PROFILE SCORE SHEET

	RUBRIC	TALLY	RAW SCORE	SEMES HRS
1	Calculus			
2	Probability/Statistics			
3	Maths Other			
4	Computer Use/Programming			
5	Physics			
6	Aerospace Engineering			
7	Mechanical Engineering			
8	Civil/Architectural Engineering			
9	Electronic/Electrical Engineering			
10	Engineering Other			
11	Physical Science, Other			
12	Biological/Agricultural/Medical Science			
13	Communications Skills			
14	Administration/Management			
15	Business Studies			
16	Education/Social/Behavioral Sciences			
17	Law			
18	Arts/Fine Arts/Humanities			
19	Military Studies			
20	Miscellaneous			

TOTAL SEMESTER HOURS __________

APPENDIX C

INSTRUCTIONS FOR USE OF PROFILE SCORE SHEET

1. You are requested to score each of the 50 transcripts herewith, using the appropriate scoring sheet for each.

2. Read the definitions of the 20 rubrics carefully. It is necessary to perceive the intention of each rubric and also those subjects specifically included or excluded.

3. For each transcript in turn:

a. Decide whether the transcript is scored in Semester hours, Credit hours or Units.

b. Study the Grading Scale. Only courses in which the candidate scores a passing grade or better are to be counted. Strike out scores relating to failures, withdrawls, incompletes, etc.

c. For each course in turn, decide on the most appropriate rubric. It may be necessary to consider department and course code to make a final decision.

d. Enter the hours (or units) earned on the tally sheet. Ignore quality or grade point scores - only hours earned are required.

e. Total the tally score for each rubric and convert to Semester hours using:

1 quarter hour = 3/4 semester hour
1 unit = 4 semester hours

f. Total the semester hours. Final scores should be of the order 120-140 semester hours for the undergraduate transcripts being examined.

CASE NUMBER: ____________

APPENDIX D

OFFICER EDUCATION PROFILE

	DEGREE	MAJOR	COLLEGE
1	BA	English	Hendrix College, Ark.
2			
3			

EDUCATION PROFILE		(Semester Hours)	
MATH	Calculus		
	Probability or Statistics	1	
	Other Mathematics	2	3
Engineering and Physics	Aerospace Engineering		
	Mechanical Engineering		
	Electrical or Electronic Engineering		
	Civil or Architectural Engineering		
	Other Engineering		
	Physics		
Other Sciences	Other Physical Sciences		
	Biological, Agricultural or Medical	6	
	Social, Behavioral, Educational or Political	19	
	Computer Programming or Use		25
Officer Managerial Studies	Basic Communication Skills	4	
	Administration or Management		
	Business Studies		
	Law		
	Military Studies		4
General Studies	Arts, Fine Arts, Humanities	71	
	Miscellaneous	4	75
Grade Point Average 3.18		TOTAL	107

APPENDIX E:

PREDICTOR VARIABLES USED TO QUANTIFY DATA IN EDUCATIONAL PROFILES

1. College Major in Agricultural, Biological or Medical Science = 1, 0 otherwise
2. College Major in Business Studies = 1, 0 otherwise
3. College Major in Psychology or Education = 1, 0 otherwise
4. College Major in Engineering = 1, 0 otherwise
5. College Major in Fine Arts = 1, 0 otherwise
6. College Major in Mathematics = 1, 0 otherwise
7. College Major in Physical Sciences = 1, 0 otherwise
8. College Major in Political Sciences = 1, 0 otherwise
9. College Major in Social Sciences = 1, 0 otherwise
10. Calculus Score
11. Probability or Statistics Score
12. Other Mathematics
13. Aerospace Engineering
14. Mechanical Engineering
15. Electrical Electronic Engineering
16. Civil or Architectural Engineering
17. Other Engineering
18. Physics
19. Other Physical Sciences
20. Biological, Agricultural or Medical Sciences
21. Social, Behavioral Educational or Political Sciences
22. Computer Programming or Use
23. Basic Communication Skills
24. Administration or Management
25. Business Studies
26. Law
27. Military Studies
28. Arts, Fine Arts, Humanities
29. Miscellaneous
30. Maths Area Score
31. Engineering and Physics Area Score
32. Other Sciences Area Score
33. Officer Managerial Studies Area Score
34. General Studies Area Score
35. Education Profile Total Score
36. Existence of Grade Point Average Score 1 if score exists, 0 otherwise
37. Grade Point Average
38. Intellectualism
39. Estheticism
40. Status
41. Pragmtatism
42. Masculinity
43. Selectivity
44. Size
45. Realistic Orientation
46. Scientific Orientation
47. Social Orientation
48. Conventional Orientation
49. Enterprising Orientation
50. Artistic Orientation

*[Predictors 10-35 are semister hours and predictors 38-50 relate to Collge Quality Scores - see "Who Goes Where to College" by Astin A.W.].

APPLICATION OF A METHOD FOR DETERMINING OFFICER GRADE REQUIREMENTS

By

Gary N. Matthews
William J. Stacy
Occupational and Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland AFB, Texas

I. INTRODUCTION

This report describes the application by Management Engineering Teams (METs) of a methodology for determining the appropriate grade requirements for officer positions. The specialties of pilots, navigators, physicians, dentists, and OGLA-exempted were excluded from consideration in this study. Essentially, the method represents a modification and extension of the Officer Grade Requirements (OGR) technology developed during the 1963-1966 time period. The objective of the recent research was to develop a methodology with which METs could determine appropriate grades for officer positions based on job content and responsibility. More specifically, the study consisted of METs collecting officer job descriptions and rating the descriptions on specified variables. Reliability and validity analyses were performed to evaluate the MET application of the technology.

II. REVIEW OF OFFICER GRADE REQUIREMENTS RESEARCH

In order to fully understand the purposes and methods of the 1974 MET study, a review of the developments in OGR research during the 1963-1966 time period is helpful as this research provided the background for the present study.

The OGR studies took place in three phases:

Phase I - Job Sampling and Policy Board Ratings. Approximately 80,000 job descriptions were completed by officers in the field. From this set, 3,575 descriptions representing all levels and types of officer jobs were selected as a criterion sample. A USAF Policy Board consisting of 22 colonels with considerable Air Force experience independently evaluated and assigned grades for each description. Exceptionally high interrater agreement ($r = .92$) was found between Policy Board members in their ratings of grades for jobs. On an additional confidence-in-rating scale, Board members expressed high confidence in their own ratings. Biases did not appear toward jobs in any command or specialty. Board members tended to agree on the specific jobs they felt were misgraded, and agreed on the direction and magnitude of the grade changes needed. The conclusion of the analyses was that each job was critically evaluated on its own merit according to its content and responsibility level.

Phase II - Development of the Policy Equation. Job evaluation factors used by the Policy Board members in assigning grades were identified and incorporated into a Policy Equation. The factors identified were: (1) special training and work experience, (2) communication skills, (3) judgment and decision-making, (4) planning, and (5) management. In addition to these five factors, other variables included in the Policy Equation were: (6) mean grade rating from raters, (7) organizational level of the job, (8) level of the job within the organization, and (9) supervisor's judgment of grade.

Phase III - Application of the Policy Equation. This equation was applied to 10,000 additional ratings of job descriptions. Using a subset of 1,750 criterion jobs, grades assigned by the Policy Equation were compared of the grades assigned the same jobs by the Policy Board using ratings collected from raters in the field. These two sets of grade scores correlated .93 and had approximately equal means and standard deviations. It was concluded that the Policy Equation had shown high stability across time and judges.

From approximately 11,000 officer job descriptions which had been collected and rated, it was possible to make projections as to how the grade structure of the Air Force might change if the Policy Equation were universally applied to determine appropriate grades for jobs.

Table 1. UDL vs OGR Grade Requirements

(Total)

	(A)	(B)	(C)
Grade	UDL	OGR	Change
Colonel	3,894	4,563	-669
LT Col	12,512	12,352	160
Major	21,809	30,776	-8,967
Lt/Capt	67,693	58,217	9,476

Results of the projections indicated that changes in grade allocations would have to be made in many officer utilization fields to bring statements of grade requirements into line with job demands. In some utilization fields, grade requirements appeared to be grossly overstated; in others they were understated. In every utilization field, some jobs were over-classified while others were under-classified by grade level.

For example, the study indicated that *too many* colonel positions were allocated to the Missile Operations, Safety, Weather, Missile Maintenance, Manpower Management, Education and Training, Information, and Legal utilization fields. At the same time, *too few* colonel positions were allocated to the Operations, Aircraft Control, Weapons Director, Scientific, Research and Development, Development Engineering, Communication-Electronics, Avionics/Munitions, Aircraft Maintenance, Civil Engineering, Transportation, Logistics, Procurement, Financial, Administrative, Intelligence, and Air Police utilization fields.

The OGR project results indicated that the Air Force could get by with considerably fewer lieutenant colonels in Pilot and Navigator cockpit positions. It should be remembered, however, that this project did not attempt to determine the appropriate pay levels for these positions, which would have demanded consideration of factors such as working conditions and risk; nor did it consider the need for the Air Force to keep rated officers in the cockpit in order to amortize the cost of their training, which is considerable.

Overall, the OGR project indicated that the Air Force was somewhat undergraded at the colonel and lieutenant colonel levels in 1964, and *considerably* undergraded at the major level.

Development of Benchmark Scales. Critical to deriving an appropriate grade for individual officer positions was the development of reliable measuring instruments for each of the five job factors in the Policy Equation. In 1966, benchmark scales were developed for each factor using job titles as shown in Table 2. For each level, job titles with high interrater agreement and low standard deviations in ratings were chosen. These scales were validated against 1,000 officer positions, and a simplified integer-weight grade equation was developed for operational application to officer positions. Again the job grades yielded by this equation duplicated those of the original Policy Board ($r = .90$).

The research methodology and instruments which have so far been reviewed formed the structure upon which the 1974 research was based.

III. THE 1974 MET STUDY

Essentially, the developmental phase of the 1974 study involved the preparation of a package of materials and instructions suitable for MET

Table 2. Example of a Benchmark Scale

FACTOR 2: SPECIAL TRAINING AND WORK EXPERIENCE: The extent to which the job requires knowledges and skills which must be acquired through special training courses or on-the-job experience. Disregard general courses given by Squadron Officer School, Command and Staff College, or War College.

LEVEL 9

Chief, Contract Pricing Branch, Hq USAF
Chief, Military Justice Division, Air Div
Space Vehicle Research Officer, Hq AF Special Weapons Center

LEVEL 8

Director, Reconnaissance & Electronic Warfare Operations, Major Air Command
Minuteman Trajectory Engineer, Aerospace Reconnaissance Technical Wg
Chief, Missile/Nuclear Safety Division, Technical Training Center

LEVEL 7

Missile Combat Crew Commander, Strategic Missile Sq
Chief, Target Intelligence Branch, Strategic Reconnaissance Wg
Chief, Maintenance Operations Div, Aerospace Test Gp

LEVEL 6

Chief, Consolidated Base Personnel Office, Bomb Wg
Flying Safety Officer, Hq Tactical Fighter Wg
Reconnaissance Aircraft Commander, Strategic Reconnaissance Sq

LEVEL 5

Co-pilot B-52, Bomb Sq
Pilot, Military Airlift Sq
Radar Evaluation Officer, Hq Major Air Command (Overseas)

LEVEL 4

Chief, Audio-Visual Center, Numbered Air Force
Electronic Warfare Officer B-52, Bomb Sq
Education-Training Officer, Major Air Command

LEVEL 3

Crypto Operations Officer, Communications Gp (Overseas)
Avionics Officer, Consolidated Aircraft Maintenance Sq
Flight Line Maintenance Officer, Organizational Maintenance Sq

LEVEL 2

Photographic Equipment Maintenance Officer, Avionics Maintenance Sq
Chief, Pay & Travel Branch, Combat Support Gp
Photographic Officer, Technical Reconnaissance Sq

LEVEL 1

Base Housing Officer, Combat Support Gp
Special Service Officer, Fighter Gp
Transportation Officer, Instrumentation Sq

personnel to collect and evaluate officer job descriptions. The field test consisted of METs obtaining job descriptions from specified job incumbents and providing factor and grade ratings on both MET-collected job descriptions and job descriptions from the previous OGR project. The evaluation phase concerned various analyses of the MET data to: (a) determine the level of a-reement among MET ratings for new and old job descriptions, (b) examine the validity of MET ratings in predicting grades established by the Air Force Policy Board for a set of jobs selected from the previous OGR project, (c) make MET-rated grade and Unit Detail Listing (UDL) grade comparisons, and (d) check for potential rating biases. If MET personnel were able to obtain suitable job descriptions and stable and valid ratings, these findings would be considered adequate to determine whether METs could accurately and operationally implement the OGR technology.

Job Samples. A main objective of the 1974 study was to determine if MET raters could produce grade ratings having a high correlation with OGR 1964 Policy Board grade ratings. In order to compare grade ratings between MET raters and the Policy Board, a common set of job descriptions rated by both groups was required. 485 non-aircrew positions which had been rated by the Policy Board and which were judged to be still current were selected.

Other objectives of the 1974 study were to determine if METs could collect accurate officer job descriptions, to determine if they could reliably rate the job descriptions on the variables developed previously in the OGR studies, and to make projections from the MET ratings to the grade structure of the Air Force. For this purpose, a second job sample of 1,687 current non-aircrew officer job descriptions was collected and rated by METs. A total of 671 MET personnel at 89 locations participated in the study.

Reliability. Interrater reliabilities of the MET ratings on the five job evaluation factors, the mean grade ratings of the MET raters, and the scores obtained from application of the equation to the ratings for the two job sets are listed in Table 3.

Table 3. Interrater Reliabilities of MET Raters

	485 Job Set	1687 Job Set
Factor 1: Special Training & Work Experience	.80	.81
Factor 2: Communication Skills	.87	.84
Factor 3: Judgment & Decision-Making	.84	.84
Factor 4: Planning	.85	.86
Factor 5: Management	.84	.87
Mean Grade Ratings of METs	.90	.93
Final Equation Scores Obtained	.94	.95

Analysis of 485 Criterion Jobs. Analysis of the rating data disclosed that grade evaluations by METs were highly correlated (r = .90) with those obtained from the Policy Board (see Table 4). Moreover, it was found that the job factors remained valid. Data revealed that for 94% of the positions, agreement was within at least two-thirds of a grade, and for only one percent of the jobs was there disagreement by more than one full grade level. This analysis indicated that MET evaluation of officer jobs could be incorporated in an operational system.

Table 4. 485 Jobs Rated by METs

Variables in OGE Equation

Variables		Validities
Factor 1: Special Training & Work Experience		.61
Factor 2: Communication Skills		.72
Factor 3: Judgment & Decision-Making		.73
Factor 4: Planning		.78
Factor 5: Management		.82
Level of Organization		.48
Level of Job Within Organization		.49
Supervisor's Judgment of Grade for Job		.79
Mean MET Grade Ratings		.86
Final Grade Evaluation Scores	r = .90	

Based on these findings, the factor ratings provided by the METs for the 485 jobs were used to construct a grade conversion table. This permits an evaluation of a job, using the grade equation to provide a score, and a conversion of the resulting score into the appropriate officer grade.

Projection of Results from the 1687 Job Sample. Grade scores on the 1,687 new jobs obtained by MET raters were determined by the conversion table. Table 5 shows a summary of these obtained scores, and compares them with UDL grade requirements and the grade of the job incumbents whose jobs were rated. The data reveal apparent inflation in the UDL at the colonel and lieutenant colonel levels and an understatement at the major level. The equation allocates a few more colonel and slightly fewer lieutenant colonel positions than held by the incumbents in the 1687 job sample. Again, the data reflect the need for more majors.

It is dangerous to generalize from 1,687 positions to the entire officer job population, especially knowing that the grade composite conversion table was constructed using only 485 cases. Yet, if one had to hazard a guess as to what the technology might yield for the total non-aircrew force, it would be reasonable to assume that the trends in Table 5 hold true.

Table 5. 1,687 New Jobs Rated by METs

Comparison of Grade Distributions

Grade Level	UMD Grade	Equation Grade	Present Grade
Colonel	121	115	108
Lt Col	307	271	275
Major	420	604	415
Lt/Capt	799	697	889
Totals:	1,647[1]	1,687	1,687

---

[1] plus 40 overages

An analysis was made to see if differences in equation allocations and UMD/UDL authorizations might be due to inflation in authorized grade distributions over time. Data presented in Table 6 suggests that there has been a significant inflation in the levels of grades authorized for non-aircrew positions since 1964.

Table 6. UDL Grade Distributions of Non-Aircrew Jobs

1964 vs 1974

	1964[1]		1974[2]	
	Freq	%	Freq	%
Colonel	3,892	5.29	5,341	8.04
Lt Col	9,356	12.72	11,588	17.44
Major	16,913	22.99	16,839	25.34
Lt/Capt	43,419	59.01	32.673	49.18
Totals:	73,580		66,441	

---

[1] from 1963 HAF 0-10 report (JR 65-10)

[2] from Feb 1974 HAF 0-8 report (AFMYC)

IV. Conclusions and Recommendations

Results of the 1974 Officer Grade Evaluation study by METs indicate that they can successfully obtain satisfactory officer job descriptions, can reliably rate officer job descriptions, and that their ratings have high validity for 1966 Policy Board ratings.

If Air Force management wishes to implement a system for evaluating officer grade requirements based on job content and responsibility, the technology described could be adopted. It is a thoroughly documented technology, and is the most defensible system available in the foreseeable future. However, additional jobs from the Criterion Board study should be rated by the METs to establish a fully defensible officer grade conversion table. Full development of this conversion table would enable the Policy Equation to be used by the Air Force in evaluating the grade for individual officer jobs. If the Air Force wishes to project grade distributions from the Policy Equation on the entire officer population of the Air Force, job evaluations on approximately 10,000 to 15,000 jobs would be required.

PHYSICAL DEMAND PROFILES FOR FOUR AIRMAN CAREER LADDERS

By

Kenneth G. Koym
Occupational and Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland AFB, Texas

I. INTRODUCTION

Since the late 1930s, various physical profiling systems have been used to medically evaluate military personnel and assign personnel to occupational specialties which do not exceed their physical abilities (Germain, Browne, and Bellows, 1953). Recently a factor "X" covering physical work capacity has been added to the system (AFM 160-1, 1975). The profile system classifies personnel according to physical abilities but no system has been developed to classify military jobs according to physical demands.

Some studies have identified physical demands in civilian jobs. In one comprehensive study, job analysts observed and interviewed job incumbents in "typical jobs" and established six physical demands for the *Dictionary of Occupational Titles* (U.S. Department of Labor, 1966). Other studies identified various physical characteristics of jobs through use of the *Position Analysis Questionnaire* (McCormick, Jeanneret, and Mecham, 1969) and the *Occupation Analysis Inventory* (Cunningham et al, 1971). These instruments contain a variety of job characteristics which could serve as source items for developing measures applicable to military jobs. No satisfactory method is yet available for evaluating the physical demands of military jobs but fairly simple and easy to administer measures of physical ability and physical proficiency of people have existed for some time (Melton, 1947; Fleishman, 1964).

In keeping with the USAF Personnel Plan objectives of making optimum man/job matches (USAFPP-1, 1971), a probe study was initiated with the broad objective of identifying, defining, measuring, and validating specific physical demands which could be used in the selection-classification process. To support the objective it was necessary to first identify and describe a limited subset of physical demands which could demonstrate significant variance in measures across career ladders and which could be objectively and economically measured in persons. Two specific objectives were involved in this investigation: (1) to determine whether the physical demands in jobs, such as dexterity, muscular coordination, and sensory discrimination, could be estimated reliably by job incumbents; and (2) to compare the physical demand profiles derived for a sample of career ladders. This paper summarizes the results of that initial effort to estimate physical demands in jobs.

II. METHOD

The Physical Demands In Jobs Survey

From a literature review, 15 physical demands were selected for use as shown in Table 1. The first ten factors were chosen with the premise that an entering airmen could be assessed on his qualifications for each demand at the point of entrance into the Air Force. Each characteristic represented a factor underlying an identifiable test (Melton, 1947; Fleishman, 1964). Five lifting demands were added to the set and each of the 15 demands was screened to insure that it was not common to a factor in the physical profiling system as described in the *Medical Examination and Medical Standards*, (AFM 160-1, 1971). Only the lifting requirements had been used by job analysts to evaluate jobs (U.S. Department of Labor, 1966). A question remained, "Could incumbents evaluate physical demands of their jobs using these factors?"

Table 1. Physical Demands in the Survey

1. Hand-Arm Movement
2. Finger Dexterity
3. Body Strength
4. Hand-Arm Strength
5. Physical Effort
6. Eye-Hand Coordination
7. Body Coordination
8. Hand-Arm Steadiness
9. Precision
10. Reaction Time
11. Lifts 1-20 Pounds
12. Lifts 21-40 Pounds
13. Lifts 41-60 Pounds
14. Lifts 61-80 Pounds
15. Lifts 81 Pounds & Over

The approach chosen to answer this question was to survey job incumbents. Each demand was given a two sentence definition. The specific wording for each of the 15 requirements is contained in the *Physical Demands in Jobs Survey* (see Appendix A).

To produce a mental set appropriate to rating the physical demands, the instructions asked the worker to (1) think of all the tasks he did in his present job and (2) to estimate how much each physical characteristic is required in the performance of his job.

In the survey, physical demands were rated using a 5-point scale, as shown in Table 2. The scale for the first 10 physical demands is in relative terms, whereas the lifting requirement scale deals with the frequency of an act. Using the latter scale, anyone who "frequently" lifted 61 to 80 pounds should have assigned a value of 4 to the 61 to 80 pound survey response block.

Table 2. Scales Used in Physical Demands Survey

Scale to Primary Physical Demands		Scale to Weight Lifting Requirement	
Value	Anchor	Value	Anchor
1	Very Little or None	1	Never
2	Small Part	2	Seldom
3	Neither Small Nor Large	3	Occasionally
4	Large Part	4	Frequently
5	Great Deal or Most	5	Almost Constantly

Selection of Career Ladders and Data Collection

Four career ladders which appeared to require different physical demands were selected for study and comparison. As shown in Table 3, they were: Avionic Instrument Systems Specialists (325X1), Instrument Repairmen (422X0), Aircraft Loadmasters (607X0/A), and Aircrew Life Support Specialists (922X0/B). Data collection was handled by mail. Completed forms were returned by 635 incumbents in the four career ladders. As indicated in Table 3, the respondents were segregated into two tenure or experience levels for subsequent analyses.

Table 3. Distribution of Rater Sample by AFSC and Experience Level

AFSC	Experience Level First Termers	 Career Personnel	Total
Avionic Instrument Systems	146	58	204
Instrument Repairman	48	24	72
Aircraft Loadmasters	78	144	222
Aircrew Life Support	78	59	137
Total			635

Data Analyses

Interrater reliability analyses were performed on the overall profile of 15 demands for each career ladder. Mean and variance values for each physical demand were computed. These values were used to make ladder-by-ladder comparisons and to prepare physical demand profiles. Additional profiles computed for first-term and career airmen were compared in an intercorrelation matrix using a within and across ladder analysis design.

III. RESULTS

Interrater Reliability Estimates

Interrater reliability estimates were derived for each career ladder using the Lindquist (1953) and Haggard (1958) components of variance technique as reported in Table 4. The R_{11} values denote the internal consistency of the physical demand profiles. The Spearman-Brown correction was used to step up the R_{11} values relative to the number of ratings per career ladder. This computation resulted in reliability values (R_{kk}) of .983 to .995. These R_{kk} values indicated that stable physical demand profiles were developed for each career ladder and less than 2 percent of variance in ratings was error variance.

Table 4. Interrater Reliability of Physical Demands Profiles

Career Ladders	K^1	R_{11}^2	R_{kk}
Avionic Instrument Systems	200	.510	.995
Instrument Repairmen	70	.511	.987
Aircraft Loadmasters	210	.238	.985
Aircrew Life Support	132	.303	.983

[1] Average number of jobs with complete sets of ratings for 15 demands.

[2] These values obtained in a rating standardization treatment used routinely in CODAP REXALL.

Physical Demand Profiles and Distributions

Profiles for the first 10 demands were plotted using the mean physical demand ratings for each of the four ladders as shown in Figure 1. Certain practical differences among the career ladders are apparent. The differences are most pronounced in eye-hand coordination, precision, and reaction time. For example, the Avionic Instrument System's profile reflects a requirement for high amounts of hand-arm movement, eye-hand coordination, and precision but relatively low amounts of body coordination and reaction time. In short, these profiles indicate that the Avionic Instrument Systems and instrument Repairman ladders demand more hand-arm movement, dexterity, eye-hand coordination, and precision than the other ladders which require strength, physical effort, and total coordination.

Data from the weight lifting requirements were excluded from the profiles due to constraints which operate when a frequency scale is used. Table 5 presents a frequency distribution of the lifting requirements. From 75 to 97 percent of the respondents indicated that they frequently or almost constantly (scale values = 4 and 5) lift up to 20 pounds, yet seldom lift 81 pounds or more. However, 35 percent of the Aircrew Life Support jobs and 18 percent of the Aircraft Loadmaster jobs require incumbents to frequently or constantly lift 81 pounds.

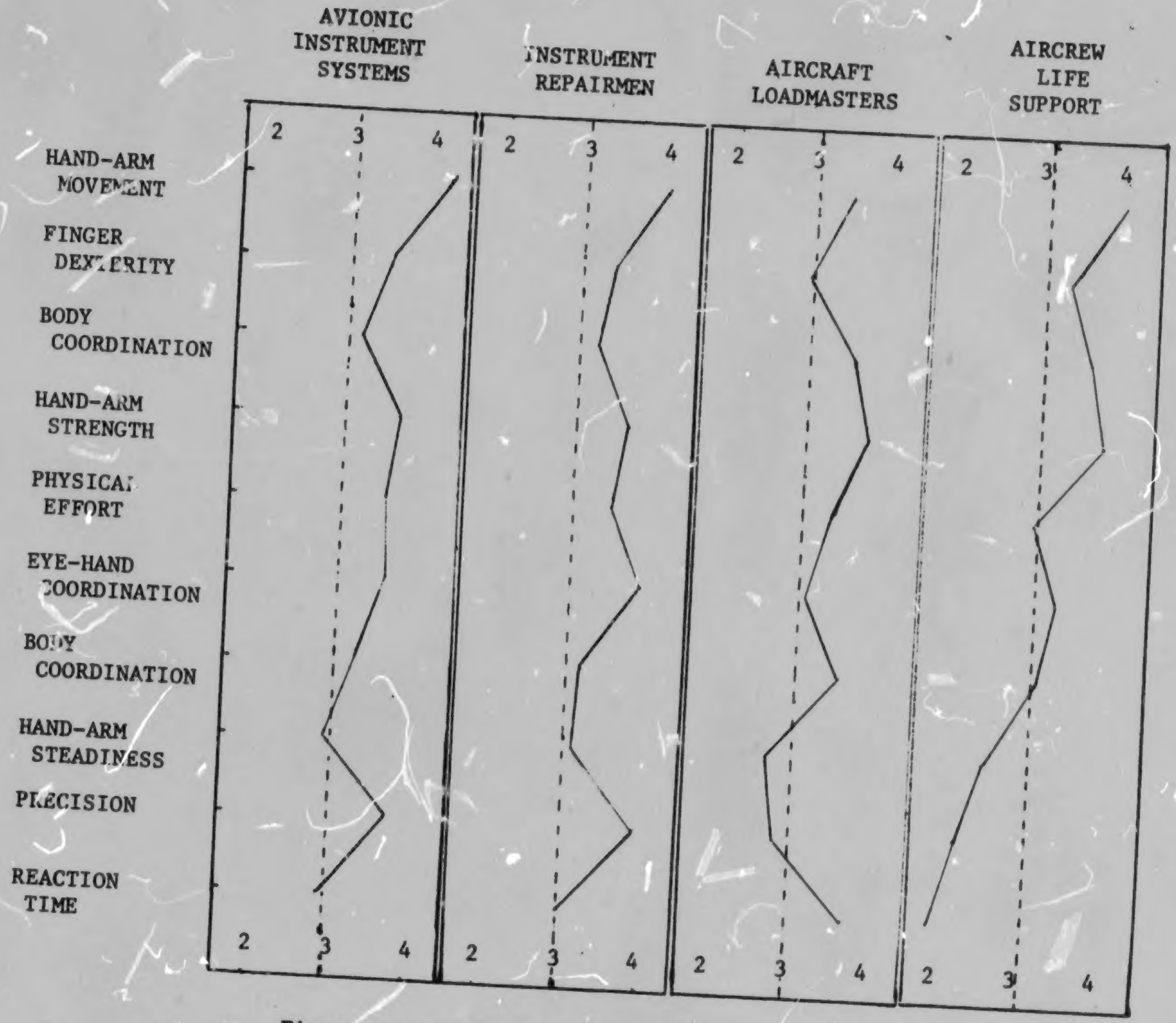


Figure 1. Physical Demand Profiles for Four Career Ladders

Table 5. Rating Distribution for Weight Lifting Categories by Career Ladder

Scale Values	% Responses To 1-20 Pounds	% Responses To 81 Pounds & Over
Avionic Instrument Systems N=189		
Seldom or Never	1.6	91.0
Occasionally	1.6	5.0
Frequently or Constantly	96.8	4.0
Instrument Repairmen N=66		
Seldom or Never	8.9	95.0
Occasionally	3.0	0.0
Frequently or Constantly	88.1	5.0
Aircraft Loadmasters N=180		
Seldom or Never	17.3	69.6
Occasionally	7.3	12.7
Frequently or Constantly	75.4	17.7
Aircrew Life Support N=122		
Seldom or Never	18.2	58.0
Occasionally	3.3	7.3
Frequently or Constantly	78.5	34.7

The mean and variance values for each physical demand were compared ladder-by-ladder, using a series of one-way analysis of variance classifications. F tests were computed to test the significance of their differences. Significant mean differences existed among the career ladders for 14 of the physical demands ($p < .01$ for 13, $p < .05$ for one demand). Hand-arm strength was not significantly different.

Tenure by Career Ladder Profile Comparisons

Means were computed for the 15 physical demand ratings for first-term airmen and careerists separately for each career ladder. These 15-observation arrays were used to compare the similarity of physical demands for the two groups within each ladder and to demands for groups outside each ladder. An intercorrelation matrix reporting the correlations within and between the four ladders is given in Table 6.

Comparing first termer and careerist relationships, profiles obtained for first termers correlate higher with those obtained for careerists within each ladder than they do with profiles for first termers in other ladders. Furthermore, the profiles for careerists correlate higher with those for first termers within their ladders than with the profiles for careerists in other ladders. The median r within ladders is .68; across ladders (for all groupings) it is .49. This finding demonstrates: (1) the ability of the survey to discriminate between career ladders better than between subsamples within ladders; and (2) the uniqueness of a pattern for a given ladder.

IV. DISCUSSION AND CONCLUSION

In conclusion, physical demands in jobs were reported using 10 factors of physical capability and proficiency for which tests have been available for some time. These requirements were determined from incumbent's, rather than job analyst's ratings. Reliable physical demand measures were obtained for four career ladders. With the exception of the hand-arm strength measure, significant physical demand differences were found between the career ladders. The most extreme differences among ladders were noted for precision and reaction time requirements.

With variations, the profiles for the Aircraft Loadmaster and Aircrew Life Support ladders demand higher amounts of physical strength (Loadmaster jobs also required fast reaction time and greater body coordination). The profiles for the Avionic Instrument Systems and Instrument Repairman ladders show higher amounts of required hand-arm movement, finger dexterity, eye-hand coordination, and precision.

For all four career ladders, a high percentage of incumbents reported a very frequent requirement for light lifting. However, an unexpectedly large number of Aircrew Life Support and Aircraft Loadmaster incumbents reported very heavy lifting requirements.

Substantially higher profile correlations were obtained between subgroups within career ladders than between airmen with similar tenure in other ladders. Results support a conclusion that career ladders have unique physical demands which can be inferred from job incumbent responses.

Table 6. Comparison of Physical Demands Profiles Within and Across AFSCs For First-Termers (F-T) and Careerists (C)

	AFSC/Group		Intercorrelations of 15 Physical Demand Ratings						
			1	2	3	4	5	6	7
1	Avionic Instrument	F-T	---						
2	Avionic Instrument	C	.947	---					
3	Instrument Repair	F-T	.891	.892	---				
4	Instrument Repair	C	.569	.563	.676	---			
5	Aircraft Loadmaster	F-T	.705	.714	.673	.397	---		
6	Aircraft Loadmaster	C	.528	.498	.480	.178	.682	---	
7	Aircrew Life Support	F-T	.439	.406	.373	.312	.490	.323	---
8	Aircrew Life Support	C	.499	.495	.447	.257	.598	.638	.783

Median r for Matrix Rearranged for Column Comparison

Within AFSCs (F-T)X(C)	Across AFSCs (F-T)X(F-T)	Across AFSCs (C)X(C)	Across AFSCs (FT)X(C)
.68	.49	.49	.49

An important task remains: That is, to validate incumbents' physical demand ratings. Future projects in this area should attempt to identify, define, measure, and validate additional physical demands. In addition, various physical demands should be evaluated as task rating factors for identifying specific career ladder requirements.

Data collected with a short physical demands survey placed in the background section of operational job inventories would permit job demands to be compared across career ladders, a capability not afforded by existing operational procedures.

REFERENCES

AF Manual 160-1. *Medical examinations and medical standards.* Washington, D.C.: Department of the Air Force, April 1971 and Change 9, May 1975.

Cunningham, J.W., Tuttle, T.C., Floyd, J.R. & Bates, J.A. *The development of the occupation analysis inventory: An "ergometric" approach to an educational problem.* Raleigh: Center for Occupational Education Research Monograph, No 6, 1971.

Fleishman, E.A. *The structure and measurement of physical fitness.* Englewood Cliffs, New Jersey: Prentice Hall, 1964.

Germain, G.L., Browne, C.G., & Bellows, R.M. Measuring men and jobs, criteria for physical analysis forms. *Personnel and Guidance Journal,* 1953, 31, 245-249.

Haggard, E.A. *Intraclass correlation and the analysis of variance.* New York: The Dryden Press, Inc., 1958.

Lindquist, E.F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin Co., 1953.

Marks, M.R. & Hook, M.E. *Development of a standard list of work requirements in airman specialties.* PRL-TR-63-19 AD-424 932. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division. August 1963.

McCormick, E.J., Jeanneret, P.R., & Mecham, R.C. *Position analysis questionnaire (PAQ).* Lafayette, Indiana: Occupational Research Center, Purdue University, 1969.

Melton, A.W. (Ed.) *Army Air Force aviation psychology program research reports, Report No. 5, apparatus tests.* Washington, D.C.: U.S. Government Printing Office, 1947.

USAFPP-1. *The USAF personnel plan: Volume I. Personnel management objectives.* Objectives 203 & 1107. Washington, D.C.: Department of the Air Force, 1971.

U.S. Department of Labor, Manpower Administration Office. *Supplement to the dictionary of occupational titles: Selected characteristics of occupations. (Physical demands, working conditions, training time).* Washington, D.C.: U.S. Government Printing Office, 1966.

APPENDIX A: PHYSICAL DEMANDS IN JOBS SURVEY

Think about all the tasks you do in your present job. Consider each of the following physical characteristics and estimate how much your job requires you to use each one in order to do the job right. For each characteristic check the one block that best describes the degree to which it plays a part in your work performance.

HAND-ARM MOVEMENT: How much of your job requires you to use closely guided movement of, or cooperation between, your arm and hand, or both arms and both hands? For example, what part of your job involves taking apart or installing medium-size components or units, or handling items in a way that requires carefully controlled movements of the hand and arm together?

Very Little Or None	Small Part	Neither Small Nor Large	Large Part	Great Deal Or Most
☐	☐	☐	☐	☐
1	2	3	4	5

(Card 01:59)

FINGER DEXTERITY: How much of your job requires you to use your fingers with quickness and skill? For example, what part of your job involves picking up and positioning or assembling small pieces, rapidly punching keys like a typewriter, or moving little items from one hand or place to another?

Very Little Or None	Small Part	Neither Small Nor Large	Large Part	Great Deal Or Most
☐	☐	☐	☐	☐
1	2	3	4	5

(60)

BODY STRENGTH: How much of your job requires you to use most of the muscles in your body to perform tasks over and over? For example, what part of your job involves withstanding muscular fatigue in the shoulders, back, and legs which results from actions like constantly driving screws with a manual tool?

Very Little Or None	Small Part	Neither Small Nor Large	Large Part	Great Deal Or Most
☐	☐	☐	☐	☐
1	2	3	4	5

(61)

HAND-ARM STRENGTH: How much of your job requires you to use your hands and arms for things like pushing, pulling or moving medium to large size objects? For example, what part of your job involves gripping tools, tightening or loosening nuts, bolts, or screws, or doing tasks that require more than just a little strength in your arms and hands?

Very Little Or None	Small Part	Neither Small Nor Large	Large Part	Great Deal Or Most
☐	☐	☐	☐	☐
1	2	3	4	5

(62)

PHYSICAL EFFORT: How much of your job requires you to use movements or positions that are tiring like working with your arms extended over your head? For example, what part of your job is done while working in cramped spaces, continuously guiding heavy objects into position, or scrambling up and down ladders, scaffolds, or stairs?

Very Little Or None	Small Part	Neither Small Nor Large	Large Part	Great Deal Or Most
☐	☐	☐	☐	☐
1	2	3	4	5

(63)

EYE-HAND COORDINATION: How much of your job requires you to use carefull coordination between your eyes and hands? For example, what part of your job involves close movements like soldering small wires, measuring small amounts accurately, or guiding very small items into holes like threading a needle?

Very Little Or None	**Small Part**	**Neither Small Nor Large**	**Large Part**	**Great Deal Or Most**
☐	☐	☐	☐	☐
1	2	3	4	5

(64)

BODY-COORDINATION: How much of your job requires you to use total body control? For example, what part of your job demands good balance and ability to move quickly and easily (not necessarily using any strength), like climbing a ladder while carrying something which prevents use of hands to control your body?

Very Little Or None	**Small Part**	**Neither Small Nor Large**	**Large Part**	**Great Deal Or Most**
☐	☐	☐	☐	☐
1	2	3	4	5

(65)

HAND-ARM STEADINESS: How much of your job requires a steady fixed positioning of the hand and arm? For example, what part of your job involves holding one position without shaking or wavering, like welding, or holding a pistol on target?

Very Little Or None	**Small Part**	**Neither Small Nor Large**	**Large Part**	**Great Deal Or Most**
☐	☐	☐	☐	☐
1	2	3	4	5

(66)

PRECISION: How much of your job requires making close or fine adjustments? For example, what part of your job demands turing knobs or dials in very small degrees, or moving levers or controls quickly and accurately, like in tuning a radio or lining up a pointer on a line scale?

Very Little Or None	**Small Part**	**Neither Small Nor Large**	**Large Part**	**Great Deal Or Most**
☐	☐	☐	☐	☐
1	2	3	4	5

(67)

REACTION TIME: How much of your job requires you to do something quickly after you get a signal by sight or by sound? For example, what part of your job involves something like flipping a switch, pushing a lever, or turning a valve immediately after hearing or seeing a signal like a buzzer or light?

Very Little Or None	**Small Part**	**Neither Small Nor Large**	**Large Part**	**Great Deal Or Most**
☐	☐	☐	☐	☐
1	2	3	4	5

(68)

Appendix A (Continued)

REQUIREMENTS FOR LIFTING:

Use the scale at the right to rate the extent to which you lift materials or objects in each of the weight ranges (shown below the scale) as a regular part of your job.

For example for Weight Category A, if your job frequently calls for lifing things that weigh between 1 and 20 pounds, then you would rate Item A as 4. If that is generally the only weight range in you job, then you would probably rate Items B, C, D, and E with a 1 (Seldom or Never).

Be sure to give each of the 5 categories its proper scale value from 1 to 5 based on what your job requires you to do.

SCALE

NEVER = 1
SELDOM = 2
OCCASIONALLY = 3
FREQUENTLY = 4
ALMOST CONSTANTLY = 5

WEIGHT CATEGORIES

A	1 - 20 pounds	☐	(69)
B	21 - 40 pounds	☐	(70)
C	41 - 60 pounds	☐	(71)
D	61 - 80 pounds	☐	(72)
E	81 pounds & over	☐	(73)

A METHOD OF DETERMINING DESIRABLE TASK EXPERIENCES FOR FIRST-LINE SUPERVISORS

By

William J. Stacy
Joe T. Hazel
Occupational & Manpower Research Division
Air Force Human Resources Laboratory (AFSC)
Lackland Air Force Base, Texas

I. INTRODUCTION

Considerable research has been accomplished on the importance of the technical abilities of first-line supervisors. Some authors have reported that when a first-line supervisor is unable to perform the tasks expected of his subordinates, work inefficiency and job dissatisfaction problems may develop for both supervisor and subordinates (Fiedler, 1967; Holsey, 1955; Katz, 1955; Likert, 1961). This report describes a method for investigating work experience deficiencies of first-line supervisors on journeyman-level tasks. Such a method may be applicable to the study of problems concerning work experience and job satisfaction.

There are reasons to suspect that certain aspects of the Air Force airman classification structure may create situations in which some personnel become first-line supervisors without experience in many of the journeyman-level tasks they are assigned to supervise. Such a situation could arise, for example, when two or more Air Force specialties merge at the 7-skill level. It may also occur in a specialty which encompasses a large number of heterogenous job types and tasks. In either case, the possibility exists that a first-line supervisor may have never performed many of the journeyman-level tasks within his own specialty.

The research reported serves a two-fold purpose. First, it describes a technique for identifying certain journeyman-level tasks which a first-line supervisor should know how to perform in order to be a satisfactory work area supervisor. In this sense those tasks are defined as "desirable" journeyman-level tasks throughout the report. The second purpose is to evaluate the extent to which supervisors of an Air Force career field have had experience in the desirable journeyman-level tasks.

The specialties and sample of job incumbents used in this study were selected from the Accounting and Finance career field.

This career field contains two specialties, General Accounting (67131/51) and Disbursement Accounting (67133/53), which merge at the 7-skill level (67170). The non-commissioned officers selected as first-line supervisors at the 7-skill level came from one or the other career specialties but are responsible for supervising airmen in either or both specialties. A practical situation was therefore presented in which the importance of supervisors having experience in desirable journeyman-level tasks could be evaluated.

II. METHOD

Three phases were involved in the approach to the present study. In order of accomplishment these were: (a) occupational analysis of two previous job inventory studies of the General Accounting (671X1) and Disbursement Accounting (671X3) career specialties (AFPTs 80-181, Jan 1970); (b) identification of desirable journeyman-level task experiences for supervisors; and (c) evaluation of supervisory career work experience in journeyman-level tasks.

Occupational Survey Analysis

Air Force occupational surveys are authorized under AFM 35-2 (1968). This program provides for collection and analysis of task data which define Air Force jobs. The methodology and analytical techniques have been reported extensively elsewhere (Archer, 1966; Archer & Fruchter, 1963; Morsh & Archer, 1967; Morsh & Christal, 1966; Morsh, Madden & Christal, 1961).

The first job inventory was a typical occupational survey of 3,246 General Accounting (671X1) and Disbursement Accounting (671X3) airmen. The inventory contained 683 significant work tasks organized into 14 duties. Each job incumbent was asked to provide personal identification and background data, check those tasks he performed in his present job, and indicate on a 7-point scale how his work time was distributed across the tasks in his job. Essentially, this analysis was used to obtain a description of work performed in the career field (e.g., job type analyses). The second occupational survey was an experimental survey designed to collect individual career work experience data on the same 683 tasks. This sample consisted of 1,261 Accounting and Finance supervisors randomly selected from the original group.

The analysis of the original job inventory which measured current job performance showed that there were eight primary job clusters or work areas in which airmen from both career specialties spent most of their work time. The eight work areas include

540 of the total 683 tasks contained in the job inventory and consist of Military Pay, Military Pay Computer, Accounts Control, Paying and Collecting, Commercial Services and Material Accounting, Travel Accounting, Civilian Pay, and Non-Appropriated Funds. These eight work areas represent the primary work performance areas of the Accounting and Finance career field. Analysis of the career work experience inventory showed that many supervisors had spent most of their career in a limited number of the work areas and thus, had experience in performing only those tasks associated with the particular work area.

Identification of Desirable Journeyman-Level Task Experiences for Supervisors

Experienced supervisors who responded to the job inventory were selected as raters to provide task importance (desirability) ratings on the journeyman-level tasks. Since the preceding job analyses had shown that many supervisors spent most of their career time within only a few of the eight primary work areas, the decision was made to limit supervisor task ratings to only those tasks in the one work area in which the rater had the most experience. Applying these restraints, 693 senior NCO's were identified, of which 405 were available to provide task importance ratings.

The 540 journeyman-level tasks were divided into eight separate booklets. Each booklet listed all the tasks which were likely to be performed in one of the eight Accounting and Finance work areas. Each supervisor rater received a booklet with a listing of the tasks in the work area in which he was the most experienced. Using a 7-point importance scale each supervisor was asked to rate every journeyman-level task in his booklet on the importance of a supervisor knowing how to perform the task to satisfactorily supervise that particular work area. The scale ranged from a rating of "1 no importance" to a rating of "7 extremely great importance."

From the supervisor ratings, task importance indices were established for each journeyman-level task by computing the arithmetic mean of all ratings given that task. Based on the 7-point importance scale, a mean of 5.00 or above was selected for identifying desirable journeyman-level tasks. There were 254 tasks with means $\geq$ 5.00 and these were subsequently used in an examination of supervisory career work experience in desirable journeyman-level tasks.

Supervisor's Career Work Experience in Desirable Journeyman-Level Tasks

Data from the 1,261 supervisors who had previously completed the career work experience occupational survey were next used to determine the extent of supervisory experience in the desirable journeyman-level tasks. Essentially, each supervisor was asked to check all the tasks performed during his career. For the present study however, only the data for the 254 tasks with a desirability mean $\geq$ 5.00 were used for the work experience analyses.

III. RESULTS AND DISCUSSION

The analyses of primary interest in this study were concerned with the task importance (desirability) ratings from the 405 senior NCOs, and the career work experience data on the desirable journeyman-level tasks for the 1,261 supervisors.

Reliability of Task Importance Ratings

One question of concern was the level of agreement among the NCO raters regarding the importance of having performed the 540 journeyman-level tasks. Assuming a high level of interrater agreement, considerable confidence could be placed on the use of task importance means to select the most desirable tasks.

Coefficients of reliability were computed using the intraclass correlation technique with adjustment for sample size (i.e., the Spearman-Brown prediction formula), described by Lindquist (1953, p. 361). The average reliability (r_{kk}) was .94 and the reliabilities for the eight work areas ranged from .87 to .95, evidencing high supervisor agreement on which journeyman-level tasks were the more important for satisfactory supervision of a particular work area. These levels of agreement suggested that the task importance ratings could be used with confidence to select the more desirable tasks.

Selection of Desirable Journeyman-Level Tasks

A total of 254 of the 540 journeyman-level tasks had a mean importance rating $\geq$ 5.00. This mean value was selected as the cutting point to designate the more desirable tasks since task ratings of 5 to 7 represented tasks of great to extremely great importance. For the 540 tasks, mean values ranged from 2.60 to 6.58. This range of mean values indicates considerable differences among tasks in terms of rated level of importance and requires a rather high cutting point for determining task desirability.

Work Experience Analyses

Through the use of several Comprehensive Occupational Data Analysis Programs (CODAP), measures of career work experience in the 254 desirable tasks were computed for various Accounting and Finance supervisory groups. Essentially, these measures consisted of the percentage of supervisors in a particular group who had checked a task as having been performed sometime during their career. These percentages for each task could then be summed and averaged across all tasks in order to determine overall career work experience for any specified group or work area. Career work experience on desirable tasks was analyzed for the following supervisory groups within the Accounting and Finance career area: (a) the total sample of 1,261 Accounting and Finance supervisors (Group 1); (b) two subsets of supervisors who could be identified as coming from either the Accounting (Group 2) or Disbursement (Group 3) career specialties. (c) three subsets of supervisors (Group 4, 5, and 6) who could be identified as currently supervising either Accounting, Disbursement, or both Accounting and Disbursement personnel. (d) eight subsets of supervisors identified as currently supervising one of the eight primary Accounting and Finance work areas.

Total Sample Work Experience Level

The average percent of members experienced in the 254 desirable tasks for the total sample of 1,261 supervisors (Group 1) was 24%. It appears, therefore, that Accounting and Finance supervisors tend to have limited work experience in the total 254 desirable tasks. This limited experience on the desirable tasks is similar to the findings for the career work experience job analyses which indicated restricted career progression across the eight primary Accounting and Finance work areas. Generally it appears that Accounting and Finance supervisors may not have experience in many of the desirable tasks needed for satisfactory supervision in the Accounting and Finance career field.

Comparison of Accounting and Disbursement Career Routes

Several analyses were made to compare career work experiences of supervisors who had reached the 7-skill level through either the Accounting (Group 2) or Disbursement (Group 3) career progression route. The average percent of members experienced in the 254 desirable tasks for supervisors from the Accounting specialty was 28% versus 22% for supervisors from the Disbursement specialty. These two mean work experience percentages indicate a rather low degree of experience in desirable tasks for both groups.

The two career specialty groups, Accounting (Group 2) and Disbursement (Group 3) were also compared on desirable tasks identified as belonging within either the Accounting or Disbursement specialties. The purpose of this analysis was to determine the extent of experience for supervisors who had come from either the Accounting or Disbursement specialties on those desirable tasks performed in both specialties. The average career experience for the Accounting (Group 2) was 48% in Accounting tasks and 20% in Disbursement tasks. The average career experience for the Disbursement (Group 3) was 42% in Disbursement tasks and 10% in Accounting tasks. These comparisons reveal that each career specialty group was significantly less experienced in the desirable tasks belonging to the other specialty than in desirable tasks of their own career specialty ($x^2 = 9.074$, df = 1, $p < .01$). While the results indicate that progression through a certain career route enhances the average level of experience in that specialty, taken together these averages suggest that Accounting and Finance supervisors may not be sufficiently experienced in many of the desirable tasks for satisfactory supervision of both the Accounting and Disbursement specialties.

Comparison of Current Supervisory Groups

Analyses similar to those used for the career route comparisons were made to compare the work experience of supervisors who were currently supervising Accounting personnel (Group 4), Disbursement personnel (Group 5), or both Accounting and Disbursement personnel (Group 6). While the average career experience of Group 4 and Group 5 are almost identical (28 percent and 27 percent respectively), the mean for supervisors of both Accounting and Disbursement personnel, Group 6 (35 percent) is larger. It appears that supervisors of both Accounting and Disbursement personnel are required to broaden their level of experience across more tasks and at a more rapid rate than supervisors of only Accounting or Disbursement personnel.

The average mean career work experiences for the Accounting (Group 4), Disbursement (Group 5), and both Accounting/Disbursement (Group 6) supervisory groups on the two subsets of tasks performed primarily in the Accounting or Disbursement specialties were studied. The average career experience for the Accounting supervisory group was 45 percent on accounting tasks and 28 percent on disbursement tasks. The average experience for the Disbursement supervisory group was 15 percent on accounting tasks and 50 percent on disbursement tasks. For the supervisors of both Accounting and Disbursement personnel (Group 6), the averages were 29 percent on accounting tasks and 55 percent on disbursement tasks.

A chi-square test of proportions indicated that these average percentages between the two subsets of tasks for Groups 4, 5, and 6 were significantly different ($x^2 = 24.51$, df = 2, $p < .001$). Comparison of the Accounting versus Disbursement supervisory groups revealed that each group was less experienced on the desirable tasks in the other specialty than on the desirable tasks in their own specialty. Supervisors of both Accounting and Disbursement personnel (Group 6) were more experienced on the disbursement tasks than the accounting tasks. Generally, these analyses suggest that many Accounting and Finance supervisors are deficient in the desirable task experiences required for satisfactory supervision in both career specialties.

Comparisons of Work Area Supervised Groups

The career work experience for 681 supervisors who were identified as currently working as a section NCOIC of one of the eight Accounting and Finance areas was analyzed. Comparisons of the mean percentages for supervisors currently supervising the eight work areas to the total sample mean (Group 1) revealed that the means for the assigned supervisors were larger than the total sample mean in all areas. A chi-square test of proportions showed the differences to be statistically significant ($x^2 = 21.05$, df = 7, $p < .001$).

The finding that supervisors in any one of the eight work areas tended to have a higher degree of experience on tasks within that specific area than supervisors from the total sample suggests a possible approach by which overall level or scope of work experience for supervisors may be increased. The work area means of the assigned supervisors tended to have the higher experience values throughout the study. If assignment within a specific area is related to a high level of experience in that area, rotation of personnel among various work areas could lead to a corresponding high level of experience across several areas. Such career work area rotation offers an assignment possibility whereby the career development of Accounting and Finance supervisors may be expanded.

Field Interviews with Accounting and Finance Personnel

In order to examine the face validity of the occupational research method concerning desirable work experiences, outcomes of the study were discussed in unstructured interviews with an Accounting and Finance officer, a deputy director, and twelve Accounting and Finance senior NCOs. The NCO group was equally divided between supervisors who were from the Accounting (671X1) and Disbursement (671X3) career ladders. Generally, members interviewed supported the findings of this study with regard to the problem of supervisory work experience deficiency in the Accounting and Finance career field.

In regard to interview comments, the consensus of opinion was that the present supervisory responsibility for the Accounting and Finance field was too broad. The unanimous recommendation was to keep the General Accounting (671X1) and Disbursement Accounting (671X3) career specialties separate through the 7-skill level.

IV. CONCLUSIONS AND RECOMMENDATIONS

This report described the development of an occupational research technique for identifying desirable journeyman tasks on which first-line supervisors (7-skill level) should have work experience. Using work experience data from a previous Accounting and Finance inventory and task importance data collected for the present study, a series of occupational analyses led to the following conclusions.

(1) From a pool of 540 journeyman-level tasks, there were 254 tasks which experienced judges rated as being highly important for supervisors to know how to perform in order to satisfactorily supervise eight Accounting and Finance work areas.

(2) Interrater reliability coefficients across all tasks and work area subsets of tasks reflected high supervisor agreement on relative task importance. Considerable confidence was demonstrated that the more desirable tasks were properly identified for subsequent work experience analyses.

(3) Several analyses of the work experiences for various supervisory groups consistently indicated that many supervisors had very limited career experience in desirable journeyman-level tasks. Many of the supervisors appeared to have spent a large portion of their career time within only a few of eight primary Accounting and Finance work areas. These findings indicated that the present Accounting and Finance personnel classification or management structure may be too broad for a supervisor to gain sufficient experience on required work activities.

(4) Field interviews with Accounting and Finance personnel tend to confirm the findings from the occupational research analyses.

Based on the preceding conclusions, the following recommendations are offered.

1. The present Accounting and Finance classification or management structure seems too broad for supervisors to gain sufficient experience across all required work activities. Since this situation seems particularly applicable to first-line supervisors, separation of the Accounting and Disbursement specialties through the 7-skill level seems warranted.[1]

2. Comparison of work experience for supervisors currently assigned to different work areas revealed that work area assignment increased level of experience on specified subsets of tasks. It appears that overall levels of work experience across a broader spectrum of work areas and tasks may be increased through a system of career work area rotation of assignments. Implementation of such a rotation assignment plan could lead to improved career development of Accounting and Finance supervisors.

3. Desirable task information and work experience levels as obtained in this report may be useful with regard to curriculum design in Accounting and Finance supervisor courses.

4. Present findings should not be considered as an actual appraisal of supervisory competency to "get-the-job done". Evaluation of supervisory competency would involve investigation of job performance or proficiency measures. Results of this study are primarily intended to demonstrate the feasibility and utility of a method for determining whether work experience deficiencies existed for first-line supervisors, and the extent of experience deficiency across various supervisory groups and functional work areas. Application of the method to other career fields where different specialties are combined at the 7-skill level appears to be an interesting and worthwhile effort for further research.

---

[1] During the time required for completion and publication of this report, Hq, USAF indicated that separation of the Accounting and Disbursement specialties through the 7-skill level was to be implemented. Present findings support this action and documentation of the recommendation appears appropriate.

REFERENCES

AF Manual 35-2. *Occupational analysis procedures for conducting occupational surveys and Air Force specialty evaluations.* Washington, D.C.: Department of the Air Force, 9 July 1968.

Archer, W. B., & Fruchter, D. A. *The construction, review, and administration of Air Force job inventories.* PRL-TDR-63-21, AD-426 755. Lackland AFB, Tex: Personnel Research Laboratory, Aerospace Medical Division, August 1963. (Contract AF 41(657)397, Psychological Research Service, Inc.)

Archer, W. B. *Computation of group job descriptions from occupational survey data.* PRL-TR-66-12, AD-653 543. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, December 1966.

Fiedler, F. E. Situational factors related to leadership effectiveness. In E. A. Fleishman (Ed.), *Studies in personnel and industrial psychology.* (26 ed.) Homewood, Ill.: The Dorsey Press, 1967, 426-436.

Halsey, G. D. *Selecting and developing first-line supervisors.* New York: Harper Brothers, 1955.

Katz, R. L. Skills of an effective administrator. *Harvard Business Review,* Vol B3, No 1, Jan-Feb 1955, 33-42.

Likert, R. *New patterns of management.* New York: McGraw-Hill, 1961.

Lindquist, E. F. *Design and Analysis of Experiments.* New York: Houghton Mifflin, 1953.

Morsh, J. E., & Christal, R. E. *Impact of the computer on job analysis in the United States Air Force.* PRL-TR-66-19, AD-656 304. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, October 1966.

Morsh, J. E., & Archer, W. B. *Procedural guide for conducting occupational surveys in the United States Air Force.* PRL-TR-67-11, AD-664 036. Lackland AFB, Tex.: Personnel Research Laboratory, Aerospace Medical Division, September 1967.

Morsh, J. E., Madden, J. M., & Christal, R. E. *Job analysis in the United States Air Force.* WADD-TR-61-113, AD-259 389. Lackland AFB, Tex.: Personnel Research Laboratory, Wright Air Development Division, February 1961.

Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

INTERSERVICE PROCEDURES FOR INSTRUCTIONAL SYSTEMS DEVELOPMENT: EXECUTIVE OVERVIEW

BY

ROBERT K. BRANSON
GAIL T. RAYNER
J. LAMARR COX
JOHN P. FURMAN

CENTER FOR EDUCATIONAL TECHNOLOGY
THE FLORIDA STATE UNIVERSITY

Interservice Procedures for Instructional Systems Development: Executive Overview

Application of the systems approach to the solution of training problems is not a new development in the military. The value of the systems approach was demonstrated in the early 1950's in military training; and in recent years, each of the services has carefully developed its own approaches to the systematic design of instruction. However, in 1973, the Interservice Committee for Instructional Systems Development (or ISD) recognized the need for a common approach to interservice training, and directed the development of an interservice ISD model.

The desirability of a uniform approach to interservice training is one obvious reason for consensus on an ISD model, but an equally important rationale was the conviction that much could be gained through combining the best systems efforts of all four services, with the most recent research from the business, industrial, and educational communities. This synthesis of ISD technology has been completed, and the resulting Instructional Systems Development Model should provide a valuable new tool not only for use in interservice training, but also for use within each of the four services.

It has long been recognized that training is a key investment in an efficient, effective fighting force. BUT it's an expensive investment requiring time, manpower, facilities, equipment, and money. The goal of ISD is to optimize the return on the military training investment; either through improving the trainee's performance on the job, or through decreasing the amount of investment required. In fact, systems designed instruction has often paid a double dividend, by both increasing performance and decreasing costs at the same time. There is convincing evidence that a systems approach to instruction is effective!

Let's take a moment to examine the results of systems engineering a sequence of courses at a Naval photographers' school: The investment in student training time decreased an average of 20%, freeing students for more advanced training or for earlier job assignment. The need for instructional personnel decreased 38%, freeing two of every five instructors for alternate assignments. One-eighth of the facilities previously occupied were released for other use. A first year equipment savings of over one-half million dollars was realized. The average per-student training cost was cut by almost one-quarter. And all of these savings were realized while maintaining or improving previous trainee performance.

The Interservice ISD model contains five phases: analysis, design, development, implementation, and control. A better view of the model can be gained through a brief look at the activities that occur in each phase.

Phase I of the ISD model includes five steps: analyzing the job, selecting tasks for training, constructing job performance measures, analyzing existing courses, and selecting the instructional setting. This is perhaps the most critical phase of ISD, because it's here that the foundation is built for the remainder of the process. Regardless of how well the following phases are carried out, if the analysis is not valid and reliable, the resulting training will not produce personnel able to perform their duties at an acceptable level.

The first step is to analyze job tasks. The output of this step is a list of the tasks that make up a particular job, with descriptions of the conditions under which the task must be performed, and the standard indicating how well the task must be performed. To find the conditions and standards of performance, the analyst must often go to locations where the job is performed. One primary source of data collection is direct observation of the task. An experienced instructor or supervisor can also provide data through expert perceptions of job tasks. A job incumbent is yet another expert source of job-task data. When available, training manuals, equipment manuals, and other published doctrine may also provide valuable assistance. The combination of the best current doctrine, and, realistic observations of field practice, provides the basis for an accurate appraisal of the tasks required in a job. After the list of job tasks performed has been developed on the basis of a small number of observations, interviews, doctrine analyses, and expert perceptions, then the task list is circulated to a large sample of experts and incumbents in the job positions, These job experts and incumbents are asked: whether the tasks are performed, how frequently they are performed, what percent of time is spent performing them, and so on. On the basis of such questions, the analyst validates the exact task content of a job.

This begins the next step of phase I which is to make decisions about which tasks, of the total actually performed, should be selected for training. The output of this step will be a list of tasks for which instruction will either be selected or developed. There are many reasons why some tasks are not trained. One example is that a task may be so simple that anyone could learn the task just by being asked to perform it by a supervisor. Other tasks are not selected because they occur so infrequently, or are done by so few people, that it's simply not practical to develop special training. Another reason why some tasks are not selected is a matter of priority in the presence of constraints in money and other resources. Selection criteria must be established to allow training resources to be placed where they will have the greatest impact on job performance. One criterion is frequency, wherein a task is rated as being performed frequently, occasionally, seldom, or never. These task ratings are then used in the selection procedure. The selection of tasks for training must be made on the basis of data collected in the field, and then interpreted by experts who are familiar with current doctrine, the setting in which the job is performed, conditions of performance, and standards of performance.

This combination of what is actually practiced in the field, and what should be practiced according to current doctrine, will result in selection of tasks for training which are as up-to-date and realistic as possible. Once the procedure is completed, a Job Performance Measure is constructed for each task which has been selected for training. A job performance measure is a valid and reliable way of measuring whether an incumbent can perform the task, under the specified conditions and to the required standards. The JPM becomes the standard against which the trainee's performance will be ultimately judged. In this sense, job performance requirements equal training requirements. After constructing job performance measures, it will be necessary to go to the field and administer the measures to be sure that they separate individuals into two groups: those who can perform the tasks, and those who cannot. This careful construction of job performance measures will ensure that the effectiveness of training can be accurately assessed, and that weaknesses in training can be quickly diagnosed.

The fourth step in Phase I is analyzing existing courses. Except in highly technical fields and in areas where hardware and procedures frequently change, much of what will be done in ISD will be the revision and improvement of existing courses. If a course is offered by another service, or by another branch of one's own service, it might be possible to avoid costly duplication of effort through adopting or revising analysis work which has already been performed. Analyzing existing courses is an exhaustive task requiring the same kind of care and attention which accompanied the first three steps of Phase I. One must look past the course title and evaluate the task analysis procedure, the selection of tasks for training, and the construction of job performance measures. Only if these systematic procedures have been followed, can the accuracy of the analysis phase be confidently accepted.

The fifth step in the analysis phase provides criteria and procedures for selecting the instructional setting. Setting means the place, method, or conditions under which training is to occur. All tasks selected for training are assigned to one or more settings. The question is to determine the most desirable combination, offering effective--yet efficient--training. The first alternative that usually comes to mind is the resident school setting. There are, however, many instances when a field setting is obviously desirable, or the task requires formal on-the-job training, or can be done using well designed job performance aids. Even if the student cannot come to the instruction, its often possible to send the instruction to the student in the form of self teaching, exportable packages.

This completes the final step of the analysis phase. The output of analysis includes the task list, job performance measures, the results of the analysis of existing courses, and the setting for each of the tasks which have been selected for training. Careful application of this type of systematic front-end analysis, helped the Army cut 600 training hours out of a Fire Control Maintenance Technicians' course--without sacrificing job performance standards. The same type of careful task analysis and task selection, coupled with extensive use of job performance aids, allowed the Navy to reduce training time for LORAN equipment repairman by 50%, while increasing trainee performance by 300%.

After completion of the Analysis effort, the Design phase begins. This phase includes: developing objectives and tests, describing student entry behavior, and determining the sequence and structure of the training. The first three steps in this phase are interdependent; however, for the purpose of describing them, they will be treated separately.

The first step in Phase II is Developing Objectives. This is where the crucial conversion of job performance measures into terminal learning objectives for training, takes place. Job tasks for which training must be provided have been identified in the first phase of the ISD model. Now it is necessary to translate these tasks into objectives for which instruction can be designed. One of the main goals of the designer at this point is to ensure a direct transfer of knowledge and skill from the training world to the job world. Most often, the job performance measure will be directly translated into a terminal learning objective which specifies: the behavior to be exhibited, the conditions under which it will be demonstrated, and the acceptable standards of performance. Statements of objectives will fall into four general categories, representing four general types of actions which students will be expected to demonstrate after training. The reason that a distinction is made among the objectives is that the training requirements for each type are often quite different.

One category of objectives is physical skills, where trainees might learn to manipulate the tools required to perform a delicate task. A second category is information, where a trainee might learn to recall the information needed to properly fill out standard forms. A third category of objectives includes mental skills like those performed by a navigator in computing the exact location of his aircraft. A final category is attitude. It may be very important that a student learns to identify unknown metals; but it is equally important that he learns a positive attitude toward safety in handling acids and other dangerous substances.

Once the objectives have been developed, the next step is to develop the tests which will measure the desired behavior specified in the objectives. There will be a direct correspondence between the objectives and the test; that is, there will be at least one test item to measure each of the principle objectives. These tests may be the performance of the actual tasks being trained, in the presence of an instructor or testing manager. The test may be a simulation of the actual task which will be rated or scored in some manner. Pencil and paper are used for testing a wide variety of objectives, just as pencil and paper are used in the direct performance of many job tasks.

Once the tests have been developed, it is time to find out what skills and knowledge the students will possess at the time when they enter the course, and whether they are at planned levels. For some training, students will enter with a very broad range of skills, abilities, experiences and prior training. It is thus important to design instruction which does not repeat that which a majority of the students already know. Nor should instruction be designed at such a high level that only a few trainees benefit from it. Since there are variations in the mental and physical abilities of trainees, and in their background experience, it is advisable to design a test of entry behavior...to determine where students stand in relation to the objectives contained in the course. Finding that a large percentage of the target population was below entry level would not mean changing the objectives, but rather extending the objectives down to a level that would include the majority of students.

The final step in the design phase is the determination of the sequence and structure of the course. Sequence and structure refers to the organization of objectives so that each step in the instruction can build on what has been previously learned. If learning one skill is dependent on first learning other skills, then that sequence is specified here. Related learning objectives are then grouped together to simplify the development of the instruction.

This concludes the steps in the design phase. The output of these four steps is a complete listing of course objectives with their test items, specification of entry behavior including data from tryouts of entry tests, and the tentative sequence in which the objectives will be taught.

An interesting experiment in the use of objectives was performed at an electronics manufacturing company where project engineers were provided with learning objectives, access to materials, and freedom to determine what they did or did not need to study. The result was a 55% decrease in training time with no loss in performance.

This approach would obviously not be suitable for all kinds of training--but it does indicate that if careful Phase I and Phase II analyses have been completed, for high quality trainees, in selected situations, one might not have to do anything more than turn the students loose with appropriate materials. In analyzing the sequencing of field wiremen's training, the Army discovered that the sttrition rate for linemen was very low during the early part of training when general principles and procedures were introduced. However, a sharp rise in attrition occured later in training when students had to learn how to climb a utility pole. By changing the course sequence and placing the pole climbing activities first, the time spent in training those students who were going to be lost through attrition, was saved. In these first two phases, a heavy investment is made in front-end analysis; but if well performed, the analysis will facilitate the completion of the ISD process, and increase the probability of an effective training product.

In the first half of this paper--the analysis phase and the design phase were described. These two steps served to define real world training needs and map out guidelines for instructional solutions.

Phase III of ISD is development. It includes specifying learning events and activities, specifying the instructional management plan and delivery system, reviewing and selecting available existing materials, and developing and validating new instruction. It is in this phase that the patient investment in front-end analysis pays off in instructional materials that accurately meet the real training needs.

The first step in phase III is to specify learning events and activities. Research has shown that for a given type of objective, there are certain general events which should occur during the instruction, if the most efficient learning is going to occur. For example, one general event required for most objectives is to provide the student with opportunities to practice what is being learned under the conditions which will occur on the job. The type of activity included in the instruction would be determined by the task that the student is learning to perform. For example, students being trained in emergency rescue procedures could practice their skills under simulated emergency conditions, or a nursing student could take actual blood pressure readings. In both of these examples, the instructional activities are designed to carry out the prescribed learning events.

After specifying appropriate learning events, and activities for each event, the next step is to specify the instructional management plan and delivery system. The management plan specifies the responsibilities required of each person associated with the course, and also specifies how instruction will be delivered to the students, and how students will be guided through that instruction. There are basically two modes in which students can be managed through instruction. One is group-mode instruction. The group may be large...or small, but common to both is the idea that all students experience the same instruction at the same pace. In some instances, the group mode of instruction is the most efficient, effective management plan available. In other instances it may be more desirable to provide a self-paced instructional mode in which individual advancement through the course is possible, or a combination of group-mode and self-paced-mode may be available. One advantage of self-pacing is that each student is managed in relation to the objectives he has or has not mastered. This approach readily allows the use of pretests for placing students in a course or even exempting them from the course. There are a variety of management plans for self-paced instruction.

- One method is peer-tutorial.
- Another is instructor-tutorial.
- Audio-tutorial is often used.
- As are the management plans using mediated instructional materials.

ISD provides the instructional developer with a broad perspective of management options, management planning decisions, along with other critical considerations, into the process of selecting the best mix of instructional media.

The media-mix incorporated in the delivery system is developed on the basis of its capacity...for transmitting or conducting the instructional activities specified in the previous step. Just because new or sophisticated equipment is available does not mean that it will be suited to the required instructional activities. For example, it would be difficult to use this computer terminal...to teach brazing techniques to a welding student. Another option to consider is the use of job performance aids like checklists, specification sheets, trouble shooting diagrams and so on, to supplement actual instruction. Finally, tentative specifications for a delivery system are subjected to cost and other resource constraints that will exist in the training setting. The delivery system thus emerges as a combination of theoretical and traditional considerations.

Developing instruction is both costly and time-consuming. It is essential to review and, when appropriate, select existing materials before deciding to commit resources to new development efforts. The instructional designer should not limit materials review to his own branch of the service--but should survey available interservice materials. Useful materials may also be available from civilian sources like universities, libraries, civic or industrial training centers, and federal government training programs. The only way to carry out a thorough review is through actual examination to determine whether the existing materials will provide the required instructional activities, and fit the management plan and delivery system. If there is no material available that can adequately help students attain mastery of specific learning objectives, or if only partially suitable materials are available, it will be necessary to develop instruction.

Before actual development begins, a careful identification must be made of production facilities and capabilities. In other words, what kind of talent can be drawn upon for the production effort? In developing instruction, as in any other complicated production process, adequate planning and management is necessary to ensure the smooth completion of the product. Once the instruction itself has been completed, attention is turned to writing student and instructor guides, and preparing other course management materials. The result of this step is a complete instructional package, composed of adequate management information plus the selected and developed materials.

Validating the materials is the only step that remains before the instruction is implemented in a formal training program. Validation is the process of trying out instruction on members of the target population and revising it until it works--that is, until the trainees achieve the objectives. A typical method of validating instruction is to have an expert examine it. However, accurately determining how students will perform is impossible on the basis of an inspection of the materials. The best basis for validating instruction is student performance. The first measure of student performance is taken in individual trials, sometimes called one-on-one trials. The instruction is administered to just one person at a time, followed by a posttest and a discussion with the student of what's unclear, what's good, and what's bad. This kind of data forms the basis on which initial revisions are made in the materials. After individual trials, group trials are planned and conducted. And again, the resulting data is used for revising the materials. The nature of the validation process is such that trials are repeated until the materials have been improved to the point where students achieve the desired performance criteria.

Validating the materials concludes Phase III. The output of Phase III is a complete management plan which specifies the delivery system, and validated learning materials containing appropriate instructional activities. Implementing creative management plans and delivery systems, offers real possibilities for significant savings in high-cost training programs. Through developing a computer managed, helicopter simulation system, the Army has eliminated 83% of the original in-flight training cost, without sacrificing student performance. The result was a projected annual savings of $7,000,000. Nineteen separate cycles were completed before developers were satisfied with basic combat training performance in the Army's experimental volunteer Army training program. Performance in the basic combat training course was measured in 16 separate testing stations. After the materials were initially developed, about 35% of the students were able to pass at least 12 of the tests on their first attempt. After five cycles, about half of the students were able to pass 12 or more of the tests on their first attempt. After ten cycles, the figure jumped to above 90%. And by the 19th cycle, over 95% of the students were passing 12 or more of the testing stations on their first attempt. Returns on further revision would have diminished drastically, because at this point there was simply little margin left for improvement.

Phase IV implements the instruction. This includes using the complete management plan, and conducting the actual course in its designated setting.

The first step is to implement the instructional management plan. The emphasis here is on ensuring... that all materials, equipment, and facilities are available and operable.

Before instruction begins, scheduling of students is completed and proper training of instructors must be ensured. When the multitude of management details have been completed...then it is time to conduct the instruction. The instruction was validated in Phase III, so here, the training is conducted:

- in the real instructional setting;
- under actual training conditions;
- with final materials;
- under the prescribed management plan.

These two steps complete the implementation phase. One produce of this phase is final documentation on course management, including: scheduling, space, resources, instructional staff, and so on. The other product is a complete cycle of instruction, including information needed for followup and improvement of the course before the next cycle.

When the Army implemented systems-engineered, self-paced instruction in wheeled vehicle maintenance, both facilities requirements and equipment requirements decreased by two-thirds.

In another example of systems-engineering, the Army implemented a peer instructional model in its field-wireman course which improved performance in some skills by as much as 300% while cutting training time more than half, from 38 days to 16 days.

The final phase of ISD is quality control. Like implementation, the control phase is a continuing effort. As long as instruction is being offered, some effort is required to monitor its effectiveness.

Internal evaluation is planned and conducted to determine whether the course meets its design goals. Information is collected not only to assess student progress, but more importantly, to assess the quality of the instruction. The process consists of collecting progress data, process data, performance data, and pertinent information from students, instructors, administrators and other expert personnel. The point of this process is to make an accurate comparison between the learning objectives originally specified, and the actual performance of students in the course. The actual learning outcomes should equal the intended learning outcomes!

After internal evaluation, one major question about the training remains unanswered. This question is whether students who have completed the training program, can do the job for which they were trained. External evaluation is conducted to answer that question. An accurate external evaluation requires...following graduates of the training program to their job assignments, and collecting data through job performance measures, questionnaires, or personal interviews. The purpose of this data collection process is to determine whether the graduate's job performance equals established job performance standards.

System revision may be required by course deficiencies discovered through internal or external evaluation, by changes in doctrine or content, or by changes in resources or student population. Revision requires as sound a data base as was used in the original analysis, design, and development of the instruction. The revision may consist of relatively minor changes to specific parts of the system, or may involve such major revision, that the designer returns to the drawing board for a review of the entire ISD process.

The output of Phase V is information on the total effectiveness of the training--information which provides the basis for accurate revision of the system.

Internal evaluation of student performance in a computer flowcharting course, indicated that low math aptitude students performed far below high math aptitude students. A major course revision, in which a self-paced management plan was incorporated, brought low aptitude students up to the point where they performed equally as well as the high aptitude students.

Analyze, design, develop, implement, control; the Interservice Training Review Program has officially approved this five-phase model. The model has been staffed throughout the Army, Navy, Air Force, and Marines--each service contributing significantly to these five phases, providing a powerful tool for interservice training.

Implementing the ISD approach offers realistic opportunities for improving the effectiveness of military training, thereby making a valuable contribution to increasing the capabilities of the Armed Forces.

INTERSERVICE PROCEDURES
FOR
INSTRUCTIONAL SYSTEMS DEVELOPMENT

IMPLICATIONS
FOR
ARMY TRAINING

WESLEY K. ROBERTS
US ARMY COMBAT ARMS TRAINING BOARD
FORT BENNING, GEORGIA 31905
SEPTEMBER, 1975

Interservice Procedures for Instructional Systems Development: Implications for Army Training

Wesley K. Roberts
US Army Combat Arms Training Board

The basis for a dynamic training program is founded in a pragmatic philosophy for curriculum development which incorporates the critical elements of training management necessary for the planning and operation of an instructional institution. A comprehensive training philosophy is difficult, at best, to ferret out in the attempt to identify an appropriate working documentation upon which an institution will base its curriculum development.

The US Army has been in the business of formally and informally training soldiers for two centuries. During this time the sophistication of weaponry has changed from simple mechanical devices into complex electronic and laser technologies. The quality of soldiers constituting the Army has risen to new levels. Today's soldiers expect not only competent leaders and challenging training, they are perceptive of disconnects between doctrine, training, and actual field practice and expect them to be rectified. The requirements for training resources have never been more demanding. A complicating obstacle to the Army training is the increasing demand for parsimonious spending of the defense budget for both weapons systems and soldier training.

Currently, the Army is undergoing an extensive effort to modernize its personnel management and training systems. This requirement has precipitated from many sources, none of which are more important than the need to train soldiers to fight outnumbered in the next war. The success of the American soldier on tomorrow's battlefield will depend significantly upon the ability of the Army's leaders to coordinate the dynamic personnel and training systems required to prepare these soldiers to "Win the first battle of the next war" today.

While there is not presently available a working document integrating the more complex features of personnel management with a pragmatic training system, a model has been developed which incorporates an implied logic for personnel management with a proven systematic methodology for curriculum development.

The Interservice Procedures for Instructional Systems Development provides a unique and well documented approach upon which the Army's training system may build its foundation. This five step model, along with its concomitant "how to" procedures, has been prepared especially for military training. An integral principle underlying the Interservice Procedures for Instructional Systems Development is inclusion of current methods used in the military services for curriculum development. The model applies a theoretical philosophy, consisting of the systems approach to empirically designed instruction, to fundamental procedures necessary in a comprehensive training system.

Analyzing Training Requirements

The first questions that arise in a systematic training model concern the analysis of curriculum requirements. In the Interservice approach this analysis begins with actual jobs. This analysis identifies tasks for jobs, verifies their performance or nonperformance both in a peacetime and combat environment. Implied in this phase is a mission analysis for collective jobs and a functional analysis for equipment as well as job tasks. A decision is then made regarding which tasks require training. A logic approach to training is one of providing instruction for those tasks that cannot otherwise be adequately performed by job aspirants and/or incumbents. Measures are then constructed to establish an acceptable indice of job performance. These job performance measures may in fact never be measurable outside of actual on-the-job occurences.

Up to this point in the Interservice model the fundamentals of a training system begin to emerge. These are the job requirements which have been identified as tasks for training. Application of blocks I.1 through I.3 of Phase I to any career specialty provides Army training managers with the heart of curriculum content. Through the application of blocks I.4 and I.5, decisions are made to include or exclude learning materials from existing training programs, to make no changes in the existing curriculum, and the appropriate instructional setting is determined. The instruction may take place in either a formal resident school or in a nonresident instruction environment. For some training, duplication of the learning materials may be necessary. That is to say, both resident and nonresident instructional settings may be demanded largely due to the overriding personnel management system. The model considers several unique instructional setting requirements.

The methodology posited in Phase I of the Interservice Procedures for Instructional Systems Development renders Army trainers with a myriad of product outcomes which include:

. job analyzes

. a list of individual tasks requiring training

. a list of collective training tasks as implied in a mission analysis

. a well defined interrelationship between individual and collective job performance tasks and the working characteristics of both equipment and the job through an implied function analysis

. job performance measures based on real-world criterion

. an identification of presently available learning materials for the conduct of required training, and

. the determination of where the training is to take place based upon well founded rationale.

A second major impact of Phase I is the cross comparison allowable in the total training system beginning at the initial state of curriculum design. Unilateral design of individual and/or collective training programs attracts major disconnects in total training system. Careful consideration during the analysis phase assists the Army training planner in establishing coherent career-long training plans which will be consistent with the requirements for collective training.

Designing Instruction

Once tasks have been selected for inclusion in a training plan and an instructional site selection has been made, it is appropriate to begin the design of instruction. In Phase II, the Interservice Instructional Systems Model provides "how to" guidance which is preliminary to learning material development.

In both individual and collective training the first step in the design of instruction is to develop objectives. These objectives are written from task statements analyzed in Phase I of the model. Learning objectives describe the behavior desired at the end of the instruction. They provide the framework for the development of test items that are used in evaluating performance in the training environment. Systematic design of instruction requires this detailed matching and orchestration of training objectives and evaluation items at the impetus of material development.

The objectives and their corresponding test items direct the description of required entry behaviors for the students. Heterogeneity in the student population may demand for a degree of remedial or advanced placement allowances in the instruction. All modern instructional systems applying state-of-the-art educational technology provides for variance in learners and maximizes individual placement into the learning materials. Such a philosophy is consistent with training individuals or units at levels which challenge their soldiering rather than repeat previously mastered instruction or abilities.

Because there is very often a logical flow of procedures leading to terminal learning objective mastery, the design of instruction necessitates this consideration. The relationships between individual and collective training objectives must be determined. Such a strategy prevents instructional redundancy and training system disconnect. The grouping of objectives allows for the clustering of related parts into instructional wholes and form the basis of what may be traditionally refered to as a lesson.

Product outcomes of a design phase in a systematic model should be predicated upon the continuation of tasks from the analysis phase. The Interservice model meets this theoretical aspect of the systems approach. Several products necessary in curriculum development are a direct result of applying the model's Phase II procedures. These are:

- learning objectives for both individual and collective training
- tests for the evaluation of learning objective mastery or nonmastery
- a description of prerequisite student behaviors which will allow for advanced placement and may require initial remedial training, and
- a plan for the structure and sequencing of both individual and collective learning objectives which are skeletal to the instructional materials.

Developing Instruction

After a thorough analysis and the completion of the design phase of an instruction system comes the actual development of the learning materials. Using the Interservice Procedures for Instructional Systems Development forces the learning materials to be developed based upon identified tasks requiring training, an extension of these tasks into learning objectives, and with the intent of teaching learners to perform tasks at the end of training which they were unable to perform at the beginning of the training.

In blocks III.1 and III.2 the learning materials are specified in terms of what the learner will do in achieving mastery of the objectives, how the instruction will be managed and how it will actually be delivered to the learner. The management plan includes such decisions as whether to have self-paced or group instruction, how the learner will be reinforced for achieving objective mastery, and the use of peer-tutors. These decisions have far-reaching impact on current Army training philosophy...that is...develop highly proficient soldiers and units with the minimum expenditure of training resources and reduce the time spent in the training base allowing for maximum time in the force. The prudent use of training management techniques is considered a vital step in the systems approach. These techniques will require the training developer to use existing materials and courses which directly apply to the learning objectives for a lesson or course. Decisions as to the learning events for the instruction and how the instruction will be managed will influence the actual preparation of the instructional packages. It is nonsensical to consider isolated development of the various forms instructional packages may take. Heretofore, resident and nonresident training have shown on a recurring basis major disconnects in the quality of instruction even though by written policy they should parallel as closely as possible. The site of the instruction is not a determinate of its quality under the aegis of the Interservice Model. Therefore, the development of instruction, both individual and collective, is always aimed at task mastery, short-circuiting this system does not adhere to sound principles nor to the philosophy frequently expressed as the etiology of our current concern in Army training often stated by General DePuy, "Winning the First Battle of the Next War."

An essential element in any modern training system is the tryout and revision of learning materials before they become officially implemented. This step provides the training manager with assurance that the materials will provide the learning experience necessary for mastery of the objectives developed for the lesson. The Interservice Model provides for validation of all instruction regardless of mode or method of presentation. This tactic has often been missing in the Army's more traditional training programs.

Through the application of Phase III techniques the Army training system stands to gain:

. coherent individual and collective training packages

. training management strategies for each unit of instruction

. quality instruction in all modes through a prior training site selection, and

. a quality assurance for each instructional unit that is an outcome of the validation process.

Implementing Instruction

Any training program will fail to achieve its potential if not properly implemented into the total training system. Materials developed under the auspices of the Interservice model are no exception.

Once the learning materials have been prepared, they must be adaptable to the training program existing in either the schools or units where they may be used. The relationship of new training materials must be clear for both soldier and trainer. The training in individual packages must compliment collective packages and vice versa. This is only possible by continuous tracking through the systematic development of instruction. When the instructional materials represent a meaningful part of the individual's own unit's training plan and when they are complete with an adequate training management plan, their successful implementation is relatively assured.

As with all training, the instruction must be conducted. The Interservice model does not place its emphasis here as other philosophies may. The conduct of instruction is determined by determinants exceeding the availability of instructors. Planning for implementation has often been an afterthought to Army training, or it has been completed by individuals wise in the ways of competent educational technology, even so this is not acceptable in today's training environment - the implementation and conduct of training must be well planned and cannot be left to other than a systematic approach.

Controlling Instruction

There are three major considerations in evaluating the performance of a training system. One is to what extent does the instruction provide a vehicle for learning in the training environment? Another, how well does the learner subsequently perform on-the-job? And, lastly, where does the system require revision?

The Interservice Procedures for Instructional Systems Development provides the evaluation or control phase (because evaluation largely controls the system) as the capstone for both individual and collective training. At the conclusion of Phase V of the model, the training manager has data evaluating the training itself, subsequent learner performance, and direction for refining and revising the system. This process allows for a cross-evaluation of individual and collective training deficiencies which summatively become the requirements of the next training cycle.

Often times, novices in educational technology failed to see the applicability of a systems approach to their particular area of training responsibility. This is almost universally true when "soft skills" provide the core of the curriculum. This logic can be empirically shaken after a single application of a comprehensive "how to" systems approach to instructional systems development (ISD). While in the past the concentration of ISD methods has found itself in the vocational training area, there is developing a solid foundation of research eivdence that so-called "soft skills" fit such a model. An example of this evidence is the design (and subsequent success) of Mager's Criterion Referenced Instruction Workshop. With time the doubters will become believers as more and more success in the ISD approach is evidenced.

Where to Apply ISD

Up to this point the rhetoric of this paper may seem unapplicable to specific programs confronting the Army training systems. A particular program that stands to benefit inordinately from the Interservice Procedures for Instructional Systems Development is the currently developing Enlisted Personnel Management System (EPMS). For purposes of brevity let's simply identify the training system components being demanded by the EPMS initiative and their logical output source under the provisions of the Interservice model:

. Soldier's Manual
  (training requirement)

. Skill Qualification Test
  (evaluation requirement)   PHASE I

. Career Training Program
  (site and time for training
  requirement

. Training for Job Objectives and

. Tests for Training
                                    PHASE II
. Description of Target Audience
  (job incumbents)

- Minimizing time in training base
    - Self-pacing
    - Extension Training
    - Streamlining Resident Courses
    - Job Performance Aids
    - Formal on-the-job Training
- Providing Training Management Schemes

PHASE III

- Fielding the Program (implementation)
- Conducting Training (in schools and units)

PHASE IV

- Evaluating Job Performance (Skill Qualification Test)
- Determining New Training Requirements (Skill Qualification Test)
- Determining System Fixes (Skill Qualification Tests)

PHASE V

While this example may appear underwhelmingly simple, it does present a logical and very practical method for achieving the results desired under the training portion of the new EPMS. The philosophy that clearly appears to permeate both the Interservice Instructional Systems Development Model and EPMS is that of defining training requirements, developing training materials, and system accountability for instruction in those job/tasks selected for training and evaluation. The adoption of a systems approach in Army training is sine qua non both by edict and prudent judgment on the part of those responsible for Army training.

Epilogue

As a matter of educational research and development priority, the US Army Combat Arms Training Board in 1973 began a project with the Center for Educational Technology of the Florida State University to design an optimal instructional systems development model for Army training. This project was designed to include a total "how to" instructional package for use in implementing the product.

This effort was very rapidly adopted by the Interservice Committee for Instructional Systems Development and therefore the project was expanded to include representation from the Air Force, Marines, and Navy.

During the two plus years of its evolution, the Interservice Instructional Systems Development Model has received both considerable praise and criticism, the latter of which has provided considerable input for revision during validation of the model. The basic principles posited by the model have been empirically established during the recent years in a number of educational projects. Because of its adaptation by the Interservice Committee for Instructional Systems Development, the model cannot specifically address itself to the requirements of any particular Armed Service. Rather, it has been written to provide a systematic model to be applied when interservice training is conducted.

Foregoing the interservice flare of its origin, the Interservice Instructional Systems Development Model provides a prudent technology for developing the instructional definitions, training materials, evaluation, and quality control mechanisms that are being required in the zeitgeist of the Army's effort to develop the training technology to "Win the First Battle of the Next War."

AN INVESTIGATION
OF THE
AIR FORCE OFFICER QUALIFYING TEST

Warren B. Shull

Presented to
the
Military Testing Association
Conference

September, 1975

Introduction

The United States Air Force Officer Qualifying Test (AFOQT) is designed to evaluate aptitudes which are important for commissioned officer performance and success. The aptitude scores derived from this test battery may also be used in counseling and classifying officers into the most suitable Air Force occupational fields and specialties. The AFOQT is given to applicants for a commission. It consists of a number of aptitude subtests arranged in several test booklets. These subtests are included in one or more of five aptitude composites. The choice of subtests included in each aptitude composite is made on the basis of studies which have shown that applicants who do well on these types of subtests also tend to do well in related types of officer training.

The Pilot Composite is a measure of some of the characteristics necessary for successful completion of pilot training, including subtests of mechanical experience, spatial information, and ability to understand and interpret information received from aircraft instruments. Applicants with high scores in this composite have considerably better chances of completing pilot training than those with low scores. The Navigator-Technical Composite is a measure of abilities to interpret dials and tables, to understand scientific and mathematical principles, and to comprehend mechanical and spatial concepts. It is designed to predict success in training courses requiring these abilities such as navigator training, communications, electronics, maintenance, engineering, and technical intelligence. The Officer Quality Composite is primarily a measure of general learning ability and officer quality. It contains measures of verbal and quantitative aptitude and reasoning ability, background knowledge relevent to world events, and an inventory of biographical material predictive of officer leadership. Applicants with high officer quality scores may be expected to do well in any technical training having appreciable academic content. The Verbal Composite is a measure of verbal skills: vocabulary, English usage, background for world events, and verbal analogies. This composite predicts success in courses such as administrative activities. The Quantitative Composite is a measure of mathematical and arithmetical reasoning ability, and ability to interpret graphs and tables. It is predictive of success in training courses in statistical services, accounting, and disbursing.[1]

The Scholastic Aptitude Test consists of two subtests: verbal and mathematics. Each is divided into two parts. The first part of the verbal subtest contains ten questions each of sentence completion, antonymes, analogies, and five reading questions on each of two articles. The second part is made up of ten reading questions of each of two articles, eight sentence completion, eight antonymes, nine analogies, and five reading questions on each of three articles. The first part of the mathematical test is 17 computational problems and 18 sufficiency data problems. The second part has 25 computational problems.[2]

The American College Testing Program examinations have four subtests: English usage, mathematical usage, social studies reading, and natural science reading. The first test consists of 75 questions covering grammar and punctuation, sentence structure, diction, and logic and organization. The mathematics test covers arithmetic and algebraic operations, arithmetic and algebraic reasoning, advanced algebra, and geometry. The social science test includes European and Ancient history, government and American history, current social issues, sociology, and economics. The natural science test covers biology, chemistry, physics, geology, astronomy, and general science.[3]

The verbal composite of the AFOQT seems to be measuring the same things as the English subtests of the ACT and SAT. Similarly, the quantitative composite of the AFOQT appears to measure the same abilities and skills as do the mathematics subtests of the ACT and SAT. The officer quality composite of the AFOQT is a weighted combination of the verbal and quantitative composites, and therefore, it might have some relationship to the composite score on the ACT. Furthermore, the pilot and navigator composites both include information on scientific and mechanical principles, which may be similar to the natural science subtest of the ACT.

Several questions may be raised. Is there a significant relationship between the composites of the AFOQT and the sub-scores on the SAT and/or ACT? What is the possibility of prediction of AFOQT scores, or the aptitudes they represent? How efficient or precise might these predictions be? The answers to these questions may permit the Air Force to revise the content and structure of the AFOQT, change its policies governing the administration and use of the test, or do both.

Methodology

This study examined the null hypotheses that there is no significant relationship between an individual's scores on the various components of the AFOQT and the subscores of the SAT or ACT. The Air Force Reserve Officers Training Corps (AFROTC) normally provides about one half of the new officers being commissioned each year and has administered the AFOQT to each. Within AFROTC those persons who have applied for a four-year scholarship represent almost 30 percent of the total number tested and is the only group for which both AFOQT and college entrance examination scores are known. The procedure of this study was to collect data on a random sample of scholarship applicants using an alphabetical composite roster of college board scores and a numerical roster of AFOQT scores. The statistical procedure used to analyze the data was the coefficient of correlation and regression analysis.

The U.S. Air Force has a long history of use of selection and classification test instruments, dating back to the second world war and the Aviation Cadet Qualifying Test and the Aircrew Classification Batteries.[4] The results of research over the past thirty years has consistently indicated that scores on the composites of the AFOQT are significantly related to the aptitudes they represent. Similar research conducted on the ACT and SAT has discovered them to be significantly related to college success. The thrust of this study was not to prove or disprove the reliability or validity of the ACT, SAT, or AFOQT. The study was conducted in the expectation of improving the administrative policies and procedures governing the use of the AFOQT.

Presentation and Analysis of the Data

This section presents data collected in the study and the results of statistical tests of the significance of correlations between AFOQT scores and SAT or ACT scores. Variable by variable analysis of the correlations between an individual's score on the Air Force Officers Qualifying Test and that person's score on either the American College Test of the Scholastic aptitude Test follow. The minimun, maximum, mean, and standard deviation for each of the eight variables recorded for the sample which took both the AFOQT and the SAT are summarized in Table 1. Similar data are presented in Table 2 for those who took the AFOQT and the ACT.

TABLE 1

Means and Standard deviations for the variables
AFOQT and SAT

Composite	Mean	SD	Min	Max
Pilot	45.2059	30.6549	1	95
Nav-Tech	38.8412	29.6016	1	95
OQC	40.3112	30.0885	1	95
Verbal	38.3529	25.3313	1	95
Quant	34.4206	28.1179	1	95
SAT-V	493.9118	95.5823	230	780
SAT-M	572.5882	105.1354	310	800
SAT-T	1065.9412	179.8082	540	1470

(N = 340)

TABLE 2

Means and Standard deviations for the variables
AFOQT and ACT

Composite	Mean	SD	Min	Max
Pilot	45.7000	29.5643	1	95
Nav-Tech	36.9250	30.0613	1	95
OQC	39.3000	30.0021	1	95
Verbal	33.7250	24.4031	1	90
Quant	32.7562	28.2114	1	95
ACT-EN	20.5500	4.2954	7	31
ACT-MA	25.3937	5.4428	3	36
ACT-SS	24.0250	5.2293	8	33
ACT-NS	26.7875	4.9556	8	34
ACT-CO	24.3125	4.0874	12	32

(N = 160)

The simple correlation coefficients between the various composites of the AFOQT, as dependent variables, and the subscores of the SAT, as independent variables, are summarized in Table 3. A similar summary for the AFOQT and the ACT is shown in Table 4.

The data compiled in this study indicated a significant correlation between dependent and independent variables in every possible combination of scores except that of the quantitative composite of the AFOQT and the partial correlation with the SAT verbal and mathematics subscores. Partial correlation refers to correlation between any two variables when the effects of the other variables have been controlled. Multiple correlation is used to indicate how much of the total variation in the dependent variable can be explained by all the independent variables acting together.[5] The verbal subscore was not significantly correlated with the quantitative score after controlling for the mathematics subscore. In addition, after controlling for the effect of the natural science subscore of the ACT, the English subscore had a negative, yet significant, correlation with the pilot composite. Due to space limitations only the most highly significant correlation between each composite and each test will be reported.

Pilot Composite

Acting together, the two subscores of the SAT explain a greater variation in the pilot composite than either subscore alone, or the total score. The pertinent statistics are shown in Table 5. Because the F ratios are significant at $p = .001$, the null hypothesis that there is no significant correlation between the pilot composite and either the SAT-V or the SAT-M subscores may be rejected. Furthermore, by using both subscores the predictability is improved over using either subscore alone.

The ACT natural science score acting alone was most predictive of the pilot composite. The computed values for the ACT-NS score and the pilot composite were $r^2 = .1673$, $a = 19.67$, $b = 2.440$, and $F = 31.75$. The regression equation can be written, Plt = (2.440) (ACT-NS) - 19.67. The standard error was 0.4331. Because the computed value of F exceeds the critical value, the null hypothesis that there is no significant correlation between the pilot composite of the AFOQT and the ACT-NS score may be rejected at $p = .001$.

TABLE 3

Simple correlation coefficients among composites of the AFOQT and subtests of the SAT

Composite	Plt	N-T	OQC	Ver	Quant	SAT-V	SAT-M	SAT-T
Pilot	1.000							
Nav-Tech	0.644	1.000						
OQC	0.411	0.698	1.000					
Verbal	0.382	0.535	0.765	1.000				
Quant	0.393	0.806	0.800	0.501	1.000			
SAT-V	0.373	0.493	0.660	0.795	0.484	1.000		
SAT-M	0.369	0.702	0.689	0.571	0.758	0.602	1.000	
SAT-T	0.417	0.673	0.754	0.757	0.702	0.884	0.904	1.000

(N = 340)

TABLE 4

Simple correlation coefficients among the composites of the AFOQT and subtests of the ACT

Composite	Plt	N-T	OQC	Ver	Qnt	EN	MA	SS	NS	CO
Pilot										
Nav-Tech	675									
OQC	385	694								
Verbal	328	574	693							
Quant	356	816	791	566						
ACT-EN	131	296	481	553	416					
ACT-MA	223	526	516	419	622	494				
ACT-SS	209	451	543	571	487	594	526			
ACT-NS	409	539	522	527	475	515	601	650		
ACT-CO	292	556	620	613	612	769	811	847	843	

(Decimals Omitted) (N = 160)

TABLE 5

Multiple and partial correlations between the pilot composite of the AFOQT and the verbal and the mathematics subscores of the SAT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
SAT-V	338	.1394	.2376	.0199	14.64	.0762
SAT-M	337	.1718	.2254	.0181	13.18	.0657

The constant was - 30.074.

Navigator-Technical Composite

The multiple and partial correlations between the navigator-technical composite and the SAT-V and SAT-T scores are shown in Table 6.

TABLE 6

Multiple and partial correlations between the navigator-technical composite of the AFOQT and the verbal and mathematics subscores of the SAT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
SAT-M	338	.4926	.6356	.0136	173.64	.1790
SAT-V	337	.5003	.1100	.0149	5.20	.0341

The constant was - 80.453.

These statistics indicate that a p = .001, the null hypothesis that there is no significant correlation between the navigator-technical composite and the SAT-M and SAT-V scores may be rejected. The data also reveal that the verbal score does not add a great deal to the explanation of the variation in the navigator-technical composite.

The significant multiple correlations of the ACT subscores with the navigator-technical composite are shown in Table 7. These data indicate that while there is a significant correlation between the navigator-technical composite and the natural science subscore, the use of the mathematics and social science subscores add significantly to the explanation of the variation in the navitagor-technical composite, but that the use of the English subscore actually reduces the ability to explain the variation. The null hypothesis that there is no significant correlation between the navigator-technical composite and the natural science, mathematics, and social science subscore of the ACT may be rejected at p = .001.

TABLE 7

Multiple and partial correlations between the navigator-technical composite of the AFOQT and the subscores of the ACT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
ACT-NS	158	.2903	.3487	.4430	18.90	2.1151
ACT-MA	157	.3543	.3165	.4865	15.58	1.7842
ACT-NS	158	.2903	.2941	.5579	10.23	1.7841
ACT-MA	157	.3543	.2943	.4538	12.83	1.6257
ACT-SS	156	.3603	.1044	.4972	1.46	0.6001
ACT-NS	158	.2903	.3082	.5608	11.11	1.8695
ACT-MA	157	.3543	.3159	.4624	14.24	1.7449
ACT-SS	156	.3603	.1468	.5317	2.52	0.8437
ACT-EN	155	.3669	-.1059	.5807	1.63	-0.7412

The constants for each of the three sets of data are, from top to bottom, -64,128, - 66.566, and -62.501.

Officer Quality Composite

The result of a partial and multiple correlation between the OQC composite score and the two SAT subscores is shown in Table 8. These data support a decision to reject, at p = .001, the null hypothesis that there was no significant correlation between the OQC composite and the two subscores of the SAT. They further indicate that using both the verbal and mathematical subscores together adds significantly to the explanation of the variation in OQC scores over the use of either subscore alone.

TABLE 8

Multiple and partial correlations between the officer quality composite of the AFOQT and the verbal and mathematics subscores of the SAT

Subscore	DOF	Partials	Beta Wt	F Ratio	Coefficient	Constant
SAT-M	338	.4741	.4571	103.93	.1308	
SAT-V	337	.5682	.3843	73.44	.1210	
						-94.3409

Multiple and partial correlations between the OQC and the ACT subscores are presented in Table 9. These data indicate that one can add significantly to the explanation of the variation in the OQC composite by using all of the ACT subscores rather than any lesser combination of subscores or any one alone. One may reject, therefore, at p = .001, the null hypothesis that there is no significant correlation between the OQC composite of the AFOQT and the four subscores of the ACT.

Verbal Composite

The result of multiple and partial correlations between the verbal composite of the AFOQT and the two SAT subtests is shown in Table 10. These data support, at p = .001, a decision to reject the null hypothesis that there was no

TABLE 9

Multiple and partial correlations between the officer quality composite of the AFOQT and the subscores of the ACT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
ACT-SS	158	.2950	.5431	.3832	66.11	3.1162
ACT-SS	158	.2950	.3758	.4281	25.36	2.1559
ACT-MA	157	.3681	.3179	.4113	18.15	1.7526
ACT-SS	158	.2950	.2988	.4749	13.03	1.7142
ACT-MA	157	.3681	.2760	.4223	12.98	1.5215
ACT-EN	156	.3848	.1668	.5655	4.24	1.1650
ACT-SS	158	.2950	.2334	.5180	6.68	1.3392
ACT-MA	157	.3681	.2238	.4505	7.50	1.2334
ACT-EN	156	.3848	.1498	.5658	3.42	1.0466
ACT-NS	155	.3968	.1583	.5463	3.08	0.9583

The four constants are -53.566, -56.999, -64.411, and -71.373, respectively.

TABLE 10

Multiple and partial correlations between the verbal composite of the AFOQT and the verbal and mathematics subscores of the SAT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
SAT-V	338	.6325	.7088	.0141	304.52	.1878
SAT-M	337	.6457	.1437	.0128	12.53	.0346

The constant was -74.252.

significant correlation between the verbal composite and the two subscores of the SAT. It also indicated that, relatively speaking, little additional explanation of the variation of the verbal composite is gained by using both the SAT-V and SAT-M instead of only the verbal score. The significant multiple correlations between the verbal composite and the ACT subscores are shown in Table 11.

TABLE 11

Multiple and partial correlations between the verbal composite of the AFOQT and the subscores of the ACT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
ACT-SS	158	.3265	.5714	0.3047	76.58	2.6664
ACT-SS	158	.3265	.3754	0.3595	23.75	1.7519
ACT-EN	157	.3970	.3300	0.4376	18.35	1.8747
ACT-SS	158	.3265	.2668	0.4082	9.30	1.2452
ACT-EN	157	.3970	.2892	0.4405	13.91	1.6431
ACT-NS	156	.4199	.2042	0.4045	6.18	1.0054

The three constants are, respectively, -30.334, -46.891, and -56.890.

These data indicate that a significant explanation of the variation of the verbal composite can be made by using the ACT social science, English, and natural science scores acting together. The null hypothesis that there is no significant correlation between the verbal score and the social science, English, and natural science subscores of the ACT may be rejected at $p = .001$.

Quantitative Composite

For the correlation between the quantitative composite and the SAT-M score, the computed values were $r^2 = .5742$, $a = -81.615$, $b = .2027$, and $F = 455.72$. The regression equation was Quant = (.2027) (SAT-M) - 81.615. The standard error was 0.0095. The computed value of F exceeded the critical value at P = .001; therefore, the null hypothesis that there was no significant correlation between the quantitative composite of the AFOQT and the mathematics score of the SAT may be rejected.

Multiple correlations were computed between the quantitative composite and the ACT subscores, and the significant results are shown in Table 12.

TABLE 12

Multiple and partial correlations between
the quantitative composite of the AFOQT
and the subscores of the ACT

Subscore	DOF	Partials	Beta Wt	Std-Err	F Ratio	Coefficient
ACT-MA	158	.3871	.6222	.3228	99.78	3.2248
ACT-MA	158	.3871	.5062	.3699	50.32	2.6238
ACT-SS	157	.4221	.2202	.3850	9.53	1.1881

The two constants were -49.133 and -62.417, respectively.

These data indicate that a significant explanation of the variation in the quantitative composite can be made by using the mathematics and social science subscores of the ACT together.

The data collected in this survey and presented above showed that there was a substantial and significant correlation between each composite score of the Air Force Officer Qualifying Test and both subscores of the Scholastic Aptitude Test, the total score of the SAT, the composite score of the American College Test, and at least one of the ACT subscores. The results of this survey are summarized below and are accompanied by recommendations.

Summary and Recommendations

This study was conducted to investigate the relationship between the Air Force Officer Qualifying Test and the two most widely used college entrance examinations, the Scholastic Aptitude Test and the American College Test. A significant relationship among the subtests of the ACT and SAT and the composites of the AFOQT would preclude the need for the Air Force to conduct the extensive testing program it does today. A substantial reduction in the scope of the testing program could result in considerable savings, in both money and man-hours, for the Air Force.

The data collected and analyzed in this survey indicated a significant correlation between each composite of the Air Force Officer Qualifying Test and both the verbal and mathematics subscores and the total score of the Scholastic Aptitude Test as well as one or more subscores and the composite score of the American College Test.

Each AFOQT composite, except the pilot, can be successfully predicted from either an individual's SAT or ACT scores. The navigator-technical composite can be predicted by either of the following equations.

$$\text{NAV} = (.0341)\ (\text{SAT-V}) + (.1790)\ (\text{SAT-M}) - 80.4533$$

$$\text{NAV} = (1.6257)\ (\text{ACT-MA}) + (.6601)\ (\text{ACT-SS}) + (1.7841)\ (\text{ACT-NS}) - 66.5661$$

The officer quality composite can be predicted by either of the following equations.

$$\text{OQC} = (.1210)\ (\text{SAT-V}) + (.1308)\ (\text{SAT-M}) - 94.3409$$

$$\text{OQC} = (1.0466)\ (\text{ACT-EN}) + (1.2334)\ (\text{ACT-MA}) + (1.3392)\ (\text{ACT-SS}) + (.9538)\ (\text{ACT-NS}) - 71.3726$$

The verbal composite can be predicted by either of these equations.

$$\text{VER} = (.1878)\ (\text{SAT-V}) + (.0346)\ (\text{SAT-M}) - 74.2519$$

$$\text{VER} = (1.6431)\ (\text{ACT-EN}) + (1.2452)\ (\text{ACT-SS}) + (1.0054)\ (\text{ACT-NS}) - 56.8897$$

The quantitative composite of the AFOQT can be predicted by either of the following equations.

$$QUANT = (.2027)\ (SAT\text{-}M) - 81.615$$

$$QUANT = (2.6238)\ (ACT\text{-}MA) + (1.1881)\ (ACT\text{-}SS) - 62.4175$$

In each case above the equation listed was the best predictor of the AFOQT composite from ACT or SAT data, as applicable. Almost seventy per cent of the variation in the composite can be explained by the appropriate equation. Conversion tables can be constructed based upon these formulae.

There is no need for the Air Force to continue to expend its resources by administering the current Air Force Officer Qualifying Test to so many thousands of people each year. A suitable substitute source of data exists in other information available to the Air Force. The Air Force can use the results of this survey in two ways. The first could be an immediate procedural change. The second could be a policy change taking longer to put into practice.

The managers of the AFROTC scholarship program could use the equations listed above to predict each applicant's AFOQT scores. Then those predicted scores should be used as a factor in the decision as to whether or not a given individual should be awarded a scholarship in place of the actual AFOQT scores used under current procedures. Pilot candidates would then be required to take the AFOQT as a precondition for activating the scholarship. The new procedures would mean that the Air Force would be administering the AFOQT to approximately 10,000 fewer persons annually. This change would save the Air Force ROTC approximately $50,000 per year.

The second recommendation for the Air Force has a more far-reaching implication. Using the equations produced by this survey, a high degree of accuracy can be achieved in predicting the navigator-technical, officer quality, verbal, and quantitative aptitudes. The Air Force should restructure the AFOQT so that only the flying aptitude portion, the pilot composite, would remain. At present the pilot portion of the AFOQT uses 48-1/2 minutes of testing time; counting time required to read instructions, that entire part takes about one hour. The total test time for the current AFOQT is six hours. A restructured test covering only flying aptitude would be approximately one-sixth as long as the present AFOQT. Five hours, therefore, would be saved each time the test was administered. Furthermore, since the new test would cover only flying aptitude, there would be no need to administer the test to anyone who was not interested in becoming a pilot. Those wishing to become navigators, missile officers, or to work in technical or non-technical fields would not be tested with the AFOQT by the Air Force; they would have the appropriate "aptitude" predicted by one of the equations presented above. This change would mean that the revised AFOQT, perhaps called the Air Force Pilot Aptitude Test, would need to be given to fewer than 2,000 people each year at an annual cost of less than $1,500.

Footnotes

[1] Air Force. Air Force Manual 35-8, Air Force Military Personnel Testing System, (Washington, DC: U.S. Government Printing Office, 1971).

[2] Martin McDonough and Alvin J. Hansen, The College Boards Examination: Complete Preparation for the Scholastic Aptitude Test (New York: Arco, 1972), pp. 2-3.

[3] American College Testing Program, Highlights of the ACT Technical Report (Iowa City: Research and Development Division, 1973), pp. 5-10.

[4] Lonnie D. Valentine, Jr., and John A Creager, Officer Selection and Classification Tests: Their Development and Use (Lackland AFB, Texas: Personnel Laboratory, 1961), p. 2.

[5] Hubert M. Blalock, Jr., Social Statistics (New York: McGraw-Hill, 1960), p. 326.

ALTERNATIVE ITEM RESPONSE WEIGHTING PROCEDURES: DEVELOPMENT AND EVALUATION[1]

by

William A. Sands

Navy Personnel Research and Development Center
San Diego, California 92152

BACKGROUND AND PURPOSE

The investigation described herein was accomplished under a program at the Navy Personnel Research and Development Center, San Diego, CA. The major purpose of this research program is the development of tools for use in selection and vocational-educational guidance for the U. S. Naval Academy midshipmen.

A rather substantial amount of research has been done on interests and their relationship to various criteria (Campbell, 1966). Using the Strong Vocational Interest Blank (SVIB), Abrahams and Neumann (1973) developed empirically-based SVIB scales designed to predict three criteria for the U. S. Naval Academy: (1) disenrollment for academic reasons; (2) disenrollment for motivational reasons; and (3) military aptitude. All three new scales evidenced significant relationships with their respective criteria in cross-validation samples.

The Naval Academy "Majors Program" was initiated in 1969 and subsequently revised. At present, academic majors are organized under three broad areas (I) Engineering-Weapons; (II) Mathematics-Science; and (III) Humanities.

Recently, there has been a marked emphasis on the importance of the technical majors included in Groups I and II. The 1974 edition of the Majors Program, published by the Academy, outlines current policy:

> The Naval Academy policy on the selection of majors is clear. Each midshipman selects a major which will meet the needs of the Navy and at the same time be

---

[1] The opinions expressed are solely those of the author and do not necessarily reflect official Department of the Navy policy. Reproduction in whole or in part is permitted for any purpose of the United States Government.

> interesting to him. The needs of the Navy take first priority and it has been determined that eighty percent of the Class shall take a technical major, i.e, Group I or II, and twenty percent may choose Group III. Hopefully, the selection of majors by the Class of 1978 will meet the 80/20 quota. If the desired distribution is not obtained by an open, free selection process, steps will be taken to adjust the distribution to meet the Navy goals. (U.S. Naval Academy, p. 9)

Pursuant to this emphasis on technical majors, Neumann and Abrahams (1974) developed a SVIB scale (E-S) designed to identify Naval Academy applicants with engineering and science interests. The class of 1973 was split into a key-development sample and a cross-validation sample. The criterion employed was dichotomous: Engineering-Science major versus "other" major. A biserial validity of .57 was obtained in the cross-validation sample. Application of the new SVIB scale to the class of 1976 yielded a biserial validity of .62. The results of this investigation were also reported in a paper presented at an Air Force symposium (Abrahams & Neumann, 1974).

The Engineering-Science scale (E-S) was developed for use in the selection of students from the pool of Academy applicants. The current operational selection composite involves a number of different predictors. The relationship between the E-S scale and the current predictors was examined and the validity of alternative composites was evaluated against various criteria; e.g., cumulative grade point average, major choice, etc. To facilitate utilization of the research findings, results have been forwarded by letter to the Dean of Admissions of the Naval Academy (Neumann, 1975). Both the Disenrollment scale and the Engineering-Science scales will be employed in computing the candidate multiple for applicants for the class of 1980 (McKee, 1975).

PROCEDURE

Instrument

The 1966 edition of the Strong Vocational Interest Blank (SVIB) for men contains 399 items dealing with occupational activities, school subjects, etc. A person taking the SVIB is asked to endorse one of three response alternatives for most of the items: "Like," "Indifferent," or "Dislike."

Samples

The U. S. Naval Academy classes of 1971, 1972, and 1973 took the SVIB during their respective plebe summers. The sample data were edited to remove persons who failed to graduate and persons who graduated with a major in the management science area. The number of persons in each specific major for each class is shown in Table 1.

Criteria

Graduates in the specific majors were aggregated into broad areas, as shown in the bottom of Table 1. Four separate problems were addressed: (1) differentiation of Engineering-Weapons majors (Group I) from all others (Groups II + III); (2) differentiation of Mathematics-Science majors (Group II) from all others (Groups I + III); (3) differentiation of Humanities majors (Group III) from all others (Groups I + II); and (4) differentiation of Engineering-Weapons majors (Group I) from Mathematics-Science majors (Group II). Previous research (Sands & McCullah, 1974) has indicated that separating Group I persons from Group II persons on the basis of their SVIB responses is more difficult than differentiating Group III majors from all others.

Item Selection

An "empirical criterion keying" approach was used to select those SVIB items having the 75 best responses for each of the four problems addressed. The proportion of high criterion group members who endorsed each of the response alternatives for each of the items was computed. The same proportion was computed for the low criterion group. Then the absolute difference between these two endorsement rates was computed. The items containing the 75 responses exhibiting the greatest absolute differences between endorsement rates were selected for subsequent weighting.

Response Weighting

Twenty alternative item response weighting methods were evaluated. In each method, weights were assigned so that high scores would be associated with the high criterion group, while the low criterion group would tend to receive lower scores.

Method #1. Unit weights were assigned to the 75 responses with the greatest absolute difference in endorsement rates. For those responses endorsed by a greater proportion of high criterion group members than by low criterion group members, a positive unit weight was assigned. Conversely, a negative unit weight was given to responses endorsed by a greater proportion of low criterion group members. Responses which were not among the best 75 received weights of zero.

Method #2. Those items having one or more responses receiving a unit weight under the first method were dimensionalized, as suggested by Campbell (1971). Each such item is considered as a continuum ranging

TABLE 1

Number of Naval Academy Graduates by Individual Major for the Classes of 1971, 1972 and 1973

Individual Major	Number of Graduates 1971	1972	1973	Total
Group I: Engineering-Weapons				
Aerospace Engineering	76	51	38	165
Electrical Engineering	22	24	22	68
General Engineering	0	0	16	16
Marine Engineering	10	13	4	27
Mechanical Engineering	66	41	37	144
Naval Architecture	15	6	10	31
Ocean Engineering	13	18	32	63
Systems Engineering	23	16	12	51
Group II: Mathematics-Science				
Applied Science	12	19	18	49
Chemistry	12	14	35	61
Mathematics	85	58	66	209
Oceanography	95	122	103	320
Operations Analysis	31	51	63	145
Physical Science	0	0	1	1
Physics	29	43	28	100
Group III: Humanities				
American Political Systems	25	49	48	122
International Security Affairs	53	54	56	163
European Studies - French	5	4	7	16
European Studies - German	15	7	3	25
European Studies - Italian	2	1	0	3
Latin American Studies - Spanish	15	12	7	34
Latin American Studies - Portuguese	0	0	2	2
Far Eastern Studies - Chinese	0	3	1	4
Soviet Studies - Russian	2	2	3	7
Economics	23	12	11	46
English	12	6	10	28
History	19	20	30	69

Summary	1971 N	1971 %	1972 N	1972 %	1973 N	1973 %	Total N	Total %
Group I	225	34	169	26	171	26	565	29
Group II	264	40	307	48	314	47	885	45
Group III	171	26	170	26	178	27	519	26

from "Like" at one extreme to "Dislike" at the opposite end. If one end of the continuum received a unit weight under the first weighting procedure, that response receives the same weight under this method and the opposite end of the response continuum is assigned a unit weight affixed with the opposite sign. If "Indifferent" was the only response for an item which received a unit weight under the first weighting procedure, the item obviously does not have the assumed underlying continuum and, therefore, all responses for the item are assigned a weight of zero under this second weighting method.

Method #3. Multiple weights were chosen after examination of the distribution of absolute differences between endorsement rates for the best 75 responses. These weights(0, ±1, ±2, ±3, ±4 and ±5) were assigned to responses according to the degree to which the two criterion groups differed. Again, positive weights were attached to responses endorsed by a greater proportion of the high criterion group than the low criterion group. Negative multiple weights were assigned to responses endorsed by a greater proportion of low than high criterion group members. The positive and negative multiple weights were assigned only to those responses receiving a unit weight under the first method. Those responses receiving a zero weight under the first method also receive a zero weight under this third method.

Method #4. The fourth weighting method is a dimensionalized version of the third method. The assumption and procedure employed in dimensionalizing was explained under the second weighting method discussed above.

Method #5. Examination of the results of other investigations suggested another set of multiple weights. These weights (0, ±3, ±4, ±5, and ±6) were assigned using essentially the same procedure as explained above for the third method. Again, weights were assigned only to those responses which received a unit weight under the first method. A weight of zero was assigned to those responses having zero weights under the first method.

Method #6. The sixth response weighting method is simply a dimensionalized version of the fifth method.

Method #7. This response weighting method assigns weights based upon the endorsement rate for the high criterion group, regardless of the low criterion group endorsement rate.

Method #8. The eighth method uses weights based upon the differences between the endorsement rates for the high and low criterion groups.

Method #9. Like the previous response weighting method, this ninth method employs the difference between the endorsement rates of the high and low criterion groups. Under the present method, this difference is squared and affixed with the sign of the nonsquared difference.

Method #10. The tenth response weighting method utilized Bayes' Theorem to estimate the probability that a person belongs to the high criterion group, given that he endorsed a particular response alternative. The desired posterior probability is a function of the conditional probabilities of endorsing the response, given membership in each of the criterion groups and the prior probabilities of belonging to each of the criterion groups.

Method #11. The posterior probability of being a member of the high criterion group, computed under the tenth method, may be considered as a proportion. Specifically, of all the persons who endorsed a particular alternative of an item, a certain proportion belong to the high criterion group. Each proportion has a standard error which is a function of the proportion itself and the sample size. This eleventh method involves inversely weighting the proportion by the standard error of the proportion.

The standard error of a proportion is influenced by the proportion itself and the sample size for the response. Specifically, the inverse of a standard error of a proportion is smallest when the proportion is 0.5 and becomes progressively larger as the proportion approaches zero or unity. This means that, for a fixed sample size, extreme proportions are weighted by a higher factor. The product of this factor (the inverse of the standard error of a proportion) and the proportion itself yields a very large weight for high proportions in comparison to the weight for low proportions, for a fixed sample size.

On the other hand, for a fixed proportion, a larger weight is assigned to those responses endorsed by a large number of persons than is given to responses made by a small number of persons. This characteristic of the eleventh weighting method reflects the belief that a proportion computed in a large sample should be more stable than would be the case for a small sample and, therefore, should receive a higher weight.

Method #12. The proportion considered in the previous two methods was weighted by a factor consisting of the ordinate on the unit normal curve divided by the standard error of the proportion. The weight assigned to a response under this twelfth method is influenced by the proportion and the sample size. As was true for the eleventh method, responses endorsed by a large number of persons receive more weight than responses based upon a small sample, when the proportion is held constant. The eleventh method modifies the original proportion by assigning larger weights to extreme proportions for a fixed sample size.

This twelfth method reverses this strategy. The original proportion is modified by assigning larger weights to proportions near 0.5 and progessively smaller weights to extreme proportions near zero or unity. The original proportion is weighted most heavily when the response distribution is equally divided by the two criterion groups.

Method #13. As mentioned above, the posterior probability of being a member of the high criterion group, given the endorsement of a particular response, is a proportion. This thirteenth method divides this proportion by its complement and performs a natural logarithmic transformation on the result.

Method #14. A two-by-two contingency table was constructed for each response alternative for each item evaluated. The rows of the tables represent the two criterion groups (high and low) while the columns represent a dichotomized response (absence versus presence).

A phi coefficient was computed for each response of each item evaluated. This coefficient is used for the fourteenth response weighting method.

Method #15. The coefficient of determination for a validity coefficient is the square of the correlation. This coefficient of determination represents the proportion of the criterion variance which is explained by the predictor variable. The phi coefficient (the validity for the dichotomized response, two criterion group problem) was squared and affixed with the sign of the nonsquared difference.

Method #16. The sixteenth method uses the phi coefficient computed for the fourteenth method and inversely weights the coefficient by the standard error of the phi coefficient.

Method #17. The magnitude of any phi coefficient is constrained by the marginal proportions of the two-by-two contingency table. In an attempt to eliminate this constraining influence on the response weights, the seventeenth method uses the ratio of the obtained phi coefficient to the maximum phi coefficient possible under the given conditions.

Method #18. The eighteenth response weighting method employs Fisher's Z coefficient, a transformation of the correlation coefficient. For two predictor categories (presence versus absence of the response) and two criterion categories (high versus low), the phi coefficient is the appropriate correlation.

Method #19. The nineteenth method uses the square of the Fisher's Z coefficient computed for the previous method. This squared coefficient is affixed with the sign of the original Z coefficient.

Method #20. The last method weights the Fisher's Z coefficient, computed for the eighteenth method, by the inverse of the standard error of the coefficient.

Scale Evaluation

The SVIB item responses for each member of the Classes of 1971, 1972, and 1973 were scored for each of the eighty scales for all midshipmen. Means and standard deviations were computed separtely for the high and low criterion groups for all scales. Point-biserial validity coefficients were computed separately for each year group, for each of the eighty scales. Finally, for each separate scale, two validities obtained in the Class of 1971 and the Class of 1972 were averaged. Each validity coefficient was transformed into a Fisher's Z coefficient, weighted by the appropriate degrees of freedom, averaged, and then converted back to a correlation. This weighted average cross-validity was used to assess the effectiveness of each of the twenty alternative item response weighting procedures. The point-biserial validities for the Class of 1973 were not used to evaluate the twenty methods, as the scales were constructed on this group.

RESULTS

Engineering-Weapons Majors Versus Other Majors

Table 2 shows the point-biserial validity coefficients for each of the twenty alternative scales for each of the three classes. These scales are designed to differentiate persons with a major in the Engineering-Weapons area from persons selecting a major in either of the other two broad academic areas.

The last column in Table 2 presents a weighted average cross-validity for each of the twenty scales. This validity is based upon the Classes of 1971 and 1972. Scales E16 and E20 demonstrated the highest average validity. A number of other scales showed closely similar validities. Specifically, eleven of the twenty scales had weighted average cross-validities within .002 of the highest average validity.

Mathematics-Science Majors Versus Other Majors

Results for the scales designed to differentiate between persons majoring in the Mathematics-Science area and persons majoring in another area are shown in Table 3. The highest weighted average validity of the twenty scales was obtained by the M04 scale. The M17 and M19 scales demonstrated weighted average cross-validities within .002 of the highest one.

TABLE 2

Effectiveness of Twenty Engineering-Science Scales

	Point-Biserial Validity Coefficients			
Scale	Key Development Sample-1973	Cross-Validation Samples 1971	Cross-Validation Samples 1972	Weighted Average Cross-Validity
E01	.419	.379	.299	.340
E02	.397	.386	.307	.348
E03	.416	.379	.304	.342
E04	.396	.384	.310	.348
E05	.419	.379	.301	.341
E06	.399	.385	.308	.347
E07	.275	.265	.245	.255
E08	.415	.380	.309	.345
E09	.410	.378	.309	.344
E10	.419	.382	.310	.347
E11	.304	.299	.265	.282
E12	.352	.343	.296	.320
E13	.364	.345	.298	.322
E14	.416	.384	.311	.348
E15	.413	.383	.309	.347
E16	.417	.384	.312	.349
E17	.408	.381	.312	.347
E18	.416	.384	.311	.348
E19	.413	.384	.311	.348
E20	.417	.384	.312	.349

TABLE 3

Effectiveness of Twenty Mathematics-Science Scales

Scale	Point-Biserial Validity Coefficients Key Development Sample-1973	Cross-Validation Samples 1971	 1972	Weighted Average Cross-Validity
M01	.399	.180	.284	.232
M02	.354	.177	.290	.234
M03	.393	.183	.293	.238
M04	.351	.186	.298	.242
M05	.399	.180	.284	.232
M06	.354	.177	.290	.234
M07	.247	.134	.212	.173
M08	.396	.173	.294	.234
M09	.389	.174	.299	.237
M10	.392	.167	.261	.214
M11	.293	.154	.229	.191
M12	.296	.152	.230	.191
M13	.286	.158	.257	.207
M14	.392	.175	.299	.237
M15	.384	.175	.302	.239
M16	.395	.175	.298	.237
M17	.370	.176	.305	.241
M18	.392	.175	.299	.237
M19	.384	.177	.302	.240
M20	.395	.175	.298	.237

Humanities Majors Versus Other Majors

Table 4 presents the point-biserial validities for each class on each of the twenty alternative scales designed to differentiate persons majoring in the area of Humanities from persons majoring in the other two broad academic areas. The highest weighted cross-validity was obtained by the H15 and H19 scales. Three other scales, H17, H18 and H20, demonstrated weighted average cross-validities within .002 of the highest one.

Engineering-Science Majors Versus Mathematics-Science Majors

The last of the four problems addressed in this study was to differentiate persons majoring in the Engineering-Science area from persons majoring in the Mathematics-Science area. The point-biserial validity coefficients for the twenty alternative item response weighting strategies for each of the three classes are presented in Table 5. The T17 scale showed the highest average cross-validity. The average validities obtained for the T04 and T06 scales were within .002 of the highest scale.

DISCUSSION AND CONCLUSIONS

Most Effective Item Response Weighting Methods

The most striking characteristic of the twenty alternative item response weighting methods is their general similarity in terms of effectiveness. For each of the four problems addressed in this study, a number of different response weighting methods have essentially the same ability to differentiate between the high and low criterion groups. A parsimonious conclusion would suggest the continued use of the common procedure of assigning positive or negative unit weights to the responses; i.e., the first method.

Least Effective Item Response Weighting Methods

Unlike the situation for the most effective method where a number of techniques are essentially equivalent, there are two methods which were consistently the least effective. Method #7 ranks as the worst method of all twenty methods. This finding holds true across all four problems. This method employs weights based upon the endorsement rate of the high criterion group, and ignores the low criterion group endorsement rate.

Method #11 was the second least effective method in each of the four problems. This method weighted the posterior probability of high criterion group membership by the inverse of the standard error.

TABLE 4

Effectiveness of Twenty Humanities Scales

	Point-Biserial Validity Coefficients			
	Key Development	Cross-Validation Samples		Weighted Average
Scale	Sample-1973	1971	1972	Cross-Validity
H01	.582	.535	.549	.542
H02	.575	.528	.543	.535
H03	.585	.534	.556	.545
H04	.581	.528	.549	.538
H05	.586	.534	.556	.545
H06	.581	.530	.550	.540
H07	.258	.271	.149	.211
H08	.581	.535	.554	.544
H09	.585	.532	.557	.544
H10	.591	.531	.565	.548
H11	.422	.391	.306	.350
H12	.505	.464	.419	.442
H13	.584	.52[illegible]	.546	.534
H14	.587	.537	.562	.549
H15	.590	.532	.572	.552
H16	.587	.536	.562	.549
H17	.585	.528	.572	.550
H18	.586	.537	.563	.550
H19	.590	.531	.572	.552
H20	.587	.537	.563	.550

TABLE 5

Effectiveness of Twenty Technical Scales

Point-Biserial Validity Coefficients

Scale	Key Development Sample-1973	Cross-Validation Samples 1971	Cross-Validation Samples 1972	Weighted Average Cross-Validity
T01	.467	.214	.213	.214
T02	.391	.253	.207	.230
T03	.473	.204	.214	.209
T04	.403	.248	.213	.231
T05	.467	.214	.214	.214
T06	.391	.254	.208	.231
T07	.212	.159	.100	.130
T08	.476	.215	.203	.209
T09	.476	.219	.199	.209
T10	.481	.223	.211	.217
T11	.275	.183	.141	.162
T12	.329	.211	.157	.184
T13	.464	.213	.205	.212
T14	.483	.221	.208	.215
T15	.488	.224	.210	.217
T16	.484	.222	.204	.213
T17	.482	.244	.222	.233
T18	.483	.222	.209	.216
T19	.488	.224	.210	.217
T20	.484	.222	.204	.213

Generalization of the Findings

This study was comprehensive in that twenty item response weighting procedures were examined in four separate problems, using three separate samples. However, the extent to which the findings can be generalized is limited by two considerations. The first limitation is that the original "best" 75 responses were selected on the basis of the greatest absolute differences in endorsement rates between the high and low criterion groups. Other procedures for selecting items to be keyed could produce results different from those reported herein.

A more important limitation is that the number of responses originally keyed was not systematically varied. For each of the four problems, those items containing the best 75 responses were chosen for keying. It is expected that one or a few of the more sophisticated differential weighting methods might evidence a distinct advantage over the simple unit weighting method if the number of keyed responses was decreased. Conversely, as key length increases, unit weighting may evidence marked advantages over the more mathematically complex methods.

Finally, the "best" method should not be determined solely on the basis of a validity coefficient. Scale test-retest reliability and scoring costs are two pertinent factors which should be included in an overall evaluation of alternative item response weighting procedures for a particular application.

REFERENCES

1. Abrahams, N. M. and Neumann, I. *The validation of the Strong Vocational Interest Blank for predicting Naval Academy disenrollment and military aptitude*. San Diego, U. S. Naval Personnel and Training Research Laboratory, February 1973. (Technical Bulletin STB 73-3).

2. Abrahams, N. M. and Neumann, I. Identification of Naval Academy applicants with engineering and science interests. *Proceedings: Psychology in the Air Force Symposium*. Colorado: U. S. Air Force Academy, April 1974.

3. Campbell, D. P. *Manual for the Strong Vocational Interest Blank*. Stanford: Stanford University Press, 1966.

4. Campbell, D. P. *Handbook for the Strong Vocational Interest Blank*. Stanford: Stanford University Press, 1971.

5. McKee, K. R. Personnel Letter to Commanding Officer, Navy Personnel Research and Development Center from Superintendent, U. S. Naval Academy, 25 August 1975.

6. Neumann, I.and Abrahams, N. M. *Identification of Naval Academy applicants with engineering and science interests*. San Diego: Navy Personnel Research and Development Center, October 1974 (Technical Report 75-7).

7. Neumann, I. Preliminary results of candidate multiple reweighting study. San Diego: Navy Personnel Research and Development Center, April 1975. (Letter to Dean of Admissions, U.S. Naval Academy).

8. Sands, W. A. and McCullah, R. D. Development and evaluation of Strong Vocational Interest Blank scales for academic major areas at the U. S. Naval Academy. Oklahoma City, Oklahoma: *Proceedings of the 16th Annual Conference of the Military Testing Association*.

9. U. S. Naval Academy. *The Majors Program*. Annapolis: Academic Dean Notice 1531 of 1 July 1974.

PREDICTION OF ARMY OFFICER PERFORMANCE

Arthur C. F. Gilbert, Ph.D.

U. S. Army Research Institute for the Behavioral and Social Sciences[1]
Arlington, Virginia 22209

Throughout the years a strong emphasis has been placed on leadership research in the U. S. Army Research Institute for the Behavioral and Social Sciences (ARI). The research has stressed defining the underlying dimensions of leadership as well as developing predictive indices in the leadership domain.

The research findings reported by Helme, Willemin, and Grafton (1971) of officer behaviors in a realistic field experimental combat setting yielded eight broad dimensions of leadership performance. Factor analysis of the Differential Officer Battery, a comprehensive set of cognitive, non-cognitive, and physical measures, defined 17 factors (Helme, Willemin, and Day, 1971). The efficacy of the battery in predicting performance on the aforementioned eight leadership factors has been reported by Helme, Willemin, and Grafton (1974).

A parallel research effort has dealt with the use of peer ratings. Downey (in press) in discussing the history of peer ratings in the Army, notes that the Ranger course was initially selected for the application of this appraisal technique because of the importance of this training for the Army leader's primary role and because of the rather extensive field exercise involved. Other research (Medland, Yates, and Downey, 1974) has dealt with the use of associate ratings in the assessment of officer potential at the senior field grade level. Downey (in press) reported on the relationship between associate evaluations obtained in Ranger training and other measures of performance in the Ranger course during the 1973 Fiscal Year.

One of the purposes of this research was to determine the validity of associate ratings and other performance measures obtained in the Ranger course in predicting performance in subsequent duty assignments. Another purpose was to assess the efficacy of a new differential performance evaluation criterion instrument.

---

[1] The views expressed in this paper are those of the author and do not necessarily reflect the views of the Army Research Institute or the Department of the Army.

PROCEDURE

Associate Ratings

A total of 313 officers who attended the Ranger Course in the second, third and fourth classes of FY 1973 were used as subjects. Associate ratings were obtained by Downey (in press) on all of these officers. He obtained three separate sets of associate ratings on these officers, one for each of the three phases of ranger training: a garrison phase at Fort Benning, a mountain camp in Georgia, and a jungle phase at Eglin Air Force Base, Florida. Data collection procedures and the relationships between associate ratings and other performance measures in the Ranger course are described in detail by Downey.

Performance Evaluations Ratings

In conjunction with a larger scale follow-up research project involving performance measures of Officer Basic Course graduates, a criterion instrument was developed. The criterion instrument, the "Performance Evaluation Form", is a modification of the officer performance evaluation procedure developed by Willemin (1965). The instrument yields a measure of overall performance as well as ratings of potential along a number of dimensions derived from the leadership research discussed and from two leadership dimensions derived from Ohio State University research (Stogdill, 1974).

In the first section of the Performance Evaluation Form, the performance appraisal required evaluation of the officer's overall performance in his principal duty assignment on a seven-point rating scale. Each step of the scale was clearly described in terms of the situational aspects involved. Thus, the high end of the scale, "7," was described as "Outstanding: Far above the requirements of the situation suggesting the highest kind of formal recognition through meritorious award, or decoration." The low end of the scale "1," was described as "Unsatisfactory: Far below the requirements of the situation, suggesting the most drastic kind of formal corrective action through reclassification, demotion, general court, or boarding out of the Army."

The second section of the Performance Evaluation Form elicited ratings concerning the areas of the officer's potential strengths for future assignment. In Table 1 a list of these areas is presented. The rater was required to rank order these functional areas in terms of the officer's potential strength for future assignment. He was then required to rate the officer's potential performance in each of these areas on the seven-step scale previously described.

The third section of the Performance Evaluation Form required a rating of the officer's potential performance under two types of leadership conditions--Combat Leadership and Technical-Managerial Leadership--and used the same seven point scale.

Data Collection

Four ratings were requested for each officer. One rating was to be accomplished by the officer's immediate superior. A second rating was to be accomplished by a superior officer other than the officer's immediate superior (but not necessarily the indorsing official) who was judged to have a knowledge of the officer's performance. The two additional ratings were to be accomplished by two close associates who were judged to be knowledgeable about the officer's performance. The three raters other than the immediate supervisor were designated by the personnel officer in consultation with the officer's immediate supervisor. Packets of ratings forms were distributed to personnel officers for 427 officers who attended the Ranger Course. Of these, usable ratings were obtained for 313 officers.

RESULTS AND DISCUSSION

The reliability estimates for the ten scales of the Performance Evaluation Form were obtained by first computing the six correlations among ratings given by the four raters for each scale. These six correlations were then averaged by using Fisher's Z-transformation (Guilford, 1965, pp. 348-349). The average correlation for each scale was adjusted by using the Spearman-Brown Prophecy Formula. The adjusted correlations for the ten scales are shown in Table 1.

The reliability estimates ranged from .56 to .63. These reliabilities are somewhat lower than those reported by Willemin (1965) who reported a reliability estimate of .73 for four raters using the same rating technique. The lower reliability estimate obtained in this research could be due to the fact that this sample of officers was relatively homogeneous with respect to performance because they constituted a more highly selected group.

The correlations among the sums of the ratings for the ten scales of the Performance Evaluation Form are shown in Table 2. These correlations were computed on 240 cases for whom complete data were available. Two regression analyses were performed to check on the differentiation of factors in the Performance Evaluation Form. Ratings on the seven scales of Part II were used to determine the ratings on the Combat Leadership Scale in the first analysis, and the same seven scales were used to determine the ratings on the Technical-Managerial Scale in the second analysis. The multiple correlation between the seven scales and the Combat Scale was .92, while the multiple correlation between the seven scales and the Technical-Managerial Scale was .89. Both of these multiple correlations are significantly beyond the .01 level.

Table 1

Reliability Estimates for the Ten Scales of the Performance Evaluation Form

Scale	Reliability Estimate
Part I	
Duty performance	.58
Part II	
Applying tactical knowledge and skills in support of combat operations	.58
Understanding the mission and clearly defining personal roles of subordinates in its accomplishment	.58
Making decisions and initiating actions under pressure	.63
Defining functional roles and duties in the process of developing subordinates to fill assignments for long-term unit effectiveness	.58
Planning and organizing manpower and materiel to meet situational requirements	.56
Motivating troops to accomplish the mission by taking into consideration their well being and morale	.56
Applying knowledge of logistics and technical matters to solve troop support problems	.60
Part III	
Combat leadership	.56
Technical-managerial leadership	.58

Table 2

Correlations[1] among the Sums of the Ratings for the Ten Scales of the Performance Evaluation Form (N=240)

Variable[2]	1	2	3	4	5	6	7	8	9	10
1	100	75	83	83	75	76	76	69	80	78
2		100	83	84	73	75	75	64	88	72
3			100	87	84	83	78	71	83	83
4				100	80	79	79	71	86	81
5					100	83	74	75	74	82
6						100	73	78	75	83
7							100	68	79	71
8								100	63	78
9									100	78
10										100

[1] Decimal points omitted from correlation coefficients.

[2] Variable numbers and discriptions:

1 = Duty performance

2 = Applying tactical knowledge and skills in support of combat operations

3 = Understanding the mission and clearly defining personal roles of subordinates in its accomplishment

4 = Making decisions and initiating actions under pressure

5 = Defining functional roles and duties in the process of developing subordinates to fill assignments for long-term unit effectiveness

6 = Planning and organizing manpower and materiel to meet situational requirements

7 = Motivating troops to accomplish the mission by taking into consideration their well being and morale

8 = Applying knowledge of logistics and technical matters to solve troop support problems

9 = Combat leadership

10 = Technical-managerial leadership

The normalized regression weights (Beta weights) for the seven scales in predicting ratings on the Combat Leadership Scale and on the Technical-Managerial Scale are shown in Table 3. Applying tactical knowledge, making decisions under pressure, and motivating troops are the three scales most related to ratings on the Combat Scale. Defining personal roles of subordinates, planning and organizing ability and decision-making are the three scales most related to ratings on the Technical-Managerial Scale.

The correlation between the sum of the four ratings on the duty performance scale and the various ratings in the Ranger Course are shown in Table 4. Of all of these, the associate ratings for each of the three phases of training and the sum of these associate ratings are most predictive of duty performance ratings. The other variables of significance in predicting duty performance ratings are the Ranger Report ratings for the three phases of training and the sum of these ratings. It should be noted at this point that the Ranger Report ratings are peer evaluations obtained within squads and are based on a forced distribution ranking procedure in a hypothethical 40-man Ranger group while associate (or ARI) evaluations are based on a nomination technique within the actual platoons. Ranger Report ratings are designed for counseling and feedback purposes while the associate ratings are designed to have greatest potential use in personnel management.

The relationship between the sum of the ratings on the different scales of the Performance Evaluation Form and the sum of the three associate ratings is shown in Table 5. Associate ratings are most predictive of ratings of potential performance in areas that demand making decisions and initiating action under pressure while they are least predictive of ratings of potential performance in the area of applying knowledge of logistics and technical matters. Also, associate ratings are more highly predictive of ratings on the Combat Leadership Scale than of ratings on the Technical-Managerial Scale.

Further analysis of the data will involve defining the relationships of the variables used in this study within career branches if sufficient sample sizes exist to obtain stable statistical estimates. Other possible analyses could center on defining the relationships that exist when the sample is segmented on the basis of certain background and demographic variables.

The methodology applied in the research will be used in a larger ARI research project that involves obtaining performance evaluation measures on approximately 5,000 officers who completed the Officer Basic Course (OBC) during Fiscal Year 1974. Of particular interest in this effort will be determining the relationship among various predictor variables and performance in the different career branches and in the Chaplain Corps.

Table 3

Beta Coefficients Obtained in Predicting
Ratings on the Combat Scale and on
the Technical-Managerial Scale

	Beta Coefficients	
Scale	Combat	Technical-Managerial
Applying tactical knowledge and skills in support of combat operations	.45015	- .06463
Understanding the mission and clearly defining personal roles of subordinates in its accomplishment	.04447	.27944
Making decisions and initiating actions under pressure	.31376	.17923
Defining functional roles and duties in the process of developing subordinates to fill assignments for long-term unit effectiveness	.01030	.15990
Planning and organizing manpower and materiel to meet situational requirements	.06663	.22306
Motivating troops to accomplish the mission by taking into consideration their well being and morale	.17987	- .04072
Applying knowledge of logistics and technical matters to solve troop support problems	- .09586	.22917

$R = .919$ $R = .893$

$R^2 = .844$ $R^2 = .797$

Table 4

Correlations between Total Duty Performance Rating and Each of the Ranger Course Ratings

Variable	Correlation
Associate Rating	
Benning Phase	.28***
Mountain Phase	.23***
Florida Phase	.27***
Total	.28***
Land Navigation Total	-.03
Physical Training Score	.06
Practical Work Examination	-.10
Patrol Grades	.06
Tactical Officer Grades	
Benning Phase	-.05
Mountain Phase	-.06
Florida Phase	.12
Total	.01
Ranger Report Grades	
Benning Phase	.16**
Mountain Phase	.16**
Florida Phase	.22***
Total	.22***
Spot Reports	.05
Special Reports	.08
Academic Status	.09
Total Ranger Performance	.01

** $p < .01$
***$p < .001$

Table 5

Correlations between Associate Ratings
and Ratings on the Ten Scales
of the Performance Evaluation Form

Scale	Correlation Form
Part I	
Duty performance	.28***
Part II	
Applying tactical knowledge and skills in support of combat operations	.31***
Understanding the mission and clearly defining personal roles of subordinates in its accomplishment	.28***
Making decisions and initiating actions under pressure	.35***
Defining functional roles and duties in the process of developing subordinates to fill assignments for long-term unit effectiveness	.28***
Planning and organizing manpower and materiel to meet situational requirements	.21***
Motivating troops to accomplish the mission by taking into consideration their well being and morale	.31***
Applying knowledge of logistics and technical matters to solve troop support problems	.19***
Part III	
Combat Leadership	.33***
Technical-managerial leadership	.24***

***$p < .001$

SUMMARY

A performance evaluation was designed to yield a measure of actual duty performance and measures of potential performance along nine dimensions of possible assignment activities. This instrument was used to obtain data on 313 officers who attended the Ranger Course in the 1973 Fiscal Year. Reliabilities estimates for the ten scales of the instrument ranged from .56 to .63. Associate ratings obtained in the Ranger Course were most predictive of actual duty performance on this instrument. The correlation between the sum of associate ratings and duty performance ratings was .28.

REFERENCES

Downey, R. G. Utilization of associate nominations in the U. S. Army officer training environment: the Ranger Course (Research Problem Review). Arlington, Va: U. S. Army Research Institute for the Behavioral and Social Sciences, in press.

Guilford, J. P. Fundamental statistics in psychology and education (Fourth Ed.). New York: McGraw-Hill, 1965.

Helme, W. H., Willemin, L. P., & Day R. W. Psychological factors measured in the differential officer battery (Technical Research Report 1173). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, July 1971 (NTIS No. AD-737 685)

Helme, W. H., Willemin, L. P., & Grafton, F. C. Dimensions of leadership in a stimulated combat situation (Technical Research Report 1172). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, July 1971. (NTIS No. AD-730 315)

Helme, W. H., Willemin, L. P. & Grafton, F. C. Prediction of officer behavior in a simulated combat situation (Research Report 1182). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, March 1974 (NTIS No. AD-779 445)

Medland, F. F., Yates, L. G. & Downey, R. G. Associate ratings of senior officer potential (Research Problem Review 74-2). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, June 1974.

Stogdill, R. M. Handbook of leadership; a survey of theory and research. New York: The Free Press, 1974.

Willemin, L. P. Criterion aspects of Army research on the prediction of officer performance (Research Study 65-6). Arlington, Va.: U. S. Army Research Institute for the Behavioral and Social Sciences, December 1965.

THE CURRENT PREDICTIVE VALIDITY OF THE FLIGHT APTITUDE SELECTION TESTS

R. L. McMullen and R. F. Eastman

U. S. Army Research Institute Field Unit
Fort Rucker, Alabama 36362

The Flight Aptitude Selection Tests (FAST), the Army's primary selection batteries for Warrant and Commissioned Officer aviators, were made operational in 1966 and have proven to be effective selection instruments. However, because of changes in initial flight training programs and the current population of aviation trainees, a revision of the FAST is underway. This study was conducted to determine how critical the need for revision is.

TABLE 1 identifies the test which comprise the two operational batteries, the Warrant Officer Candidate battery and the Officer battery.

The FAST battery encompasses four content areas: (1) Personality and Leadership, (2) Spatial Aptitude, (3) Mechanical Aptitude and Knowledge, and (4) Aviation Information. In general, the Self-Description materials obtain their validity by predicting preflight success, while the cognitive items predict flying and academic performance.

In addition to the FAST battery, a cutting score of 110 on the General Technical Aptitude Area (GT) of the ACB is to screen applicants for flight training. This makes failure for academic reasons a minor cause of attrition in training.

The pre- and post-FAST Flight Training success rates shown in TABLE 2 indicate the reduction in attrition which occurred after implementation of the batteries.

TABLE 1

COMPOSITION OF FAST BATTERIES

TEST	OFFICER	WARRANT OFFICER
BIOGRAPHICAL INFORMATION	X	
MECHANICAL PRINCIPLES	X	
FLIGHT ORIENTATION	X	
AVIATION INFORMATION	X	X
HELICOPTER INFORMATION	X	X
MECHANICAL INFORMATION	X	X
MECHANICAL FUNCTIONS	X	X
VISUALIZATION OF MANEUVERS	X	X
INSTRUMENT COMPREHENSION	X	X
COMPLEX MOVEMENTS	X	X
STICK AND RUDDER ORIENTATION	X	X
SELF-DESCRIPTION		X

TABLE 2

SUCCESS RATE IN FLIGHT TRAINING

OFFICERS	
PRIOR TO IMPLEMENTATION FAST	75%
CURRENT	90%
WO CANDIDATES	
PRIOR TO IMPLEMENTATION FAST	50%
CURRENT	65%

METHOD

The grades and training dispositions of 557 Initial Entry Rotary Wing (IERW) students, representing approximately the last two-thirds of the student input (classes 74-18 through 74-50), were provided by the Army Aviation Center. FAST answer sheet files from the years 1971-1974 were searched and the FAST scores of 264 students were identified. FAST scores for less than 50% of the population were obtained because 60% of the 373 officer trainees were ROTC and do not take the FAST-OB. Additionally, a randomly selected small percentage of FAST answer sheets are not available in ARI files.

Predictor and Criterion Variables

FAST-OB and WOCB composite scores were the predictor variables of interest. IERW grades and course dispositions were jointly employed as a criterion measure. Trainees were categorized as attritees if eliminated for any reason other than administrative or medical reasons, which were clearly not effects of the training program. All attritees were assigned a failing grade of 68, slightly less than one standard deviation below the minimum passing score, for purposes of analysis. This practice of assigning failing grades has been demonstrated to be an unbiased method of including failures in a validity analysis while retaining the linearity of regression of criterion grades on a valid predictor (Maier, 1968).

RESULTS

The means and standard deviations of FAST scores attained by aviator trainees during the original validation studies are presented in TABLE 3 with the test scores of current WOC and officer trainees (Kaplan, 1965; Rosenberg, Martinek and Anderson, 1959). It is apparent that the test scores of current WOC trainees are generally higher than those of the original validation groups. A possible cause of this effect may be the higher cutting scores currently required of WOCs.

TABLE 4 contains the means and standard deviations of FAST scores for successful and attrited trainees. Successful officer trainees obtained significantly higher FAST composite scores than did attrited officers ($p < .05$). The score difference between successful and attrited WOCs was not statistically significant. However, the more restrictive cutting score imposed on the sample may have attenuated the effect. The difference in the attrition rate presented in TABLE 2 and those appearing in TABLE 4 may represent chance fluctations, or a response to the changing relative proportions of WOCs and officers in the program (from 2:1 in favor of WOCs during the VN period to the present ratio of about 1:1).

The means and standard deviations of flight training grades for successful WOCs and officers are shown in TABLE 5. The successful officers attained a nonsignificantly higher mean final grade than did successful WOCs. The standard deviations for both groups of successful students fall between 2 and 3 points, indicating a rather narrow distribution of passing scores.

TABLE 3

MEAN AND STANDARD DEVIATION OF WOC AND OFFICER TRAINEES' FAST SCORES

TEST	ORIGINAL VALIDATION		CURRENT VALUES			
			WOCB		OB	
	MEAN	S.D.	MEAN	S.D.	MEAN	S.D.
BIOGRAPHICAL INFORMATION	23.7	6.3	NA	NA	69.6	12.2
SELF DESCRIPTION	71.9	8.6	81.4	6.2	NA	NA
MECHANICAL PRINCIPLES	15.0	6.6	NA	NA	12.7	6.3
FLIGHT ORIENTATION	30.5	17.5	NA	NA	32.2	16.6
INSTRUMENT COMPREHENSION	15.6	8.4	21.8	7.2	20.1	7.9
MECHANICAL INFORMATION	15.9	6.9	16.3	6.1	11.9	5.8
COMPLEX MOVEMENTS	28.8	8.9	26.6	9.6	27.2	12.2
VISUALIZATION OF MANEUVERS	15.1	7.1	20.0	5.4	19.1	6.5
HELICOPTER KNOWLEDGE	8.7	4.3	12.3	4.2	10.9	3.9
STICK AND RUDDER ORIENTATION	22.1	9.7	29.7	6.9	26.4	8.8
AVIATION INFORMATION	10.2	7.6	11.3	6.0	9.0	5.7
MECHANICAL FUNCTIONS	12.2	7.6	18.2	6.6	15.1	7.4

TABLE 4

MEANS, STANDARD DEVIATIONS OF FAST SCORES BY TRAINING DISPOSITION AND % ATTRITION WITHIN SAMPLE

FAST COMPOSITE SCORES

	WOC		OFFICER	
	SUCCESSFUL	ATTRITED	SUCCESSFUL	ATTRITED
$\overline{X}$	342.6	333.7	278	244
s	28.0	28.3	58.5	64.8
N	79	34	127	24
% ATTRITION	30.0		15.9	

TABLE 5

MEANS AND STANDARD DEVIATIONS OF FLIGHT TRAINING GRADES FOR SUCCESSFUL WOC AND OFFICER TRAINEES

	WOCs	OFFICERS
$\overline{X}$	82.7	85 4
s	2.3	2.9
N	79	127

TABLE 6 contains the means and standard deviations of the FAST scores and final grades for all WOC and officer trainees. Here again, the attrited trainees were assigned final grades of 68, slightly less than one standard deviation below the minimum passing score. It is apparent from TABLES 5 and 6 that officers attain higher final flight training grades than WOCs. This may be due, in part, to the military development WOCs receive concurrently with flight and academic training.

The original validities for the FAST batteries in TABLE 7 were calculated separately for fixed-wing and rotary-wing composites. The Army no longer has an initial entry fixed-wing program and operational practice now is to use the overall composite scores to select IERW trainees. When the current validities and original validities of the FAST WOCB and OB compositles are compared, it appears that the predictive validity of the WOC battery has attenuated somewhat since its implementation. (The current validities have been corrected for restriction in range resulting from the FAST cutting scores.)

Separate validity coefficients were computed for setbacks (trainees who repeat a portion of the program) and nonsetbacks (TABLE 8). The contrast between the officers and WOCs, analyzed in this manner, is striking. The WOC setbacks and nonsetbacks had essentially identical FAST scores and flight training grades, while the validity coefficient for nonsetbacks was significantly greater (.58) than that for setbacks (.03). For the officers, a quite different relationship was found. The nonsetbacks had significantly higher FAST and flight training grades than setbacks ($t = 2.83$, $p < .01$; $t = 2.98$, $p < .01$) while officer setbacks demonstrated a nonsignificantly higher validity than nonsetbacks.

TABLE 6

MEANS AND STANDARD DEVIATIONS OF FAST SCORES AND FINAL FLIGHT TRAINING GRADES OF WOCs AND OFFICERS

	WOC		OFFICERS	
	FAST-WOCB COMPOSITE SCORES	FLIGHT TRAINING GRADES	FAST-OB COMPOSITE SCORES	FLIGHT TRAINING GRADES
$\bar{X}$	339.7	78.4	273.2	82.8
s	28.6	7.0	58.7	6.7
N	113		151	

TABLE 7

ORIGINAL AND CURRENT PREDICTIVE VALIDITIES OF FLIGHT APTITUDE SELECTION TESTS

ORIGINAL VALIDITY ESTIMATES

FAST FIXED WING	WOCB ROTARY WING	FAST FIXED WING	OB ROTARY WING
.46	.48	.39	.42

CURRENT PREDICTIVE VALIDITY

FAST WOCB	FAST OB
.38	.44

TABLE 8

MEANS, STANDARD DEVIATIONS AND VALIDITIES OF SETBACKS AND NONSETBACKS

	WOCS				OFFICERS			
	SETBACKS		NONSETBACKS		SETBACKS		NONSETBACKS	
	FAST	GRADES	FAST	GRADES	FAST	GRADES	FAST	GRADES
$\bar{X}$	339.7	78.1	339.9	78.6	250.2	78.1	284.2	85.1
s	29.8	6.2	28.1	7.6	49.5	8.1	60.9	4.3
N	47		65		52		99	
r	.03		.58		.47		.35	

CONCLUSIONS

1. The FAST scores obtained by current WOC trainees are higher than the scores of trainees involved in the original validation studies. The inconsistent results for the officer group suggests that this may be due to the present high cutting score for WOCs.

2. In general, officers achieve higher flight training grades than WOCs, even though the FAST-OB cutting score is very low, about 7th percentile, while the FAST-WOCB has a cutting score near the 50th percentile. Several factors contributing to this disparity include: (a) military development training required of WOCs, (b) superior academic preparation of officers, (c) a halo effect enjoyed by commissioned officers, and (d) the severe consequences of failure to an officer's career. The validity coefficient obtained for the officer sample suggests that a more stringent cutting score will reduce attrition among officer trainees.

3. The predictive validity of the WOCB has attenuated somewhat when compared with the original validation studies; however, validities obtained in this study indicate that the present battery remains a useful selection instrument until a projected revision of the battery is available for operational use.

4. The data for setbacks and nonsetbacks showed marked differences between the officer and WOC samples. The WOC setbacks and nonsetbacks had highly similar FAST scores and flight grades, while the officers displayed a more predictable pattern, i.e., higher test scores and grades for the nonsetbacks. The lack of a correlation between FAST scores and flight training grades for WOC setbacks is difficult

to interpret when compared with the high validity coefficient obtained by nonsetbacks. A possible explanation of the disparity between these groups may be the influence of a random effect military development grades have on final flight training grades of WOCs.

REFERENCES

Gulliksen, H. Theory of Mental Tests. New York: Wiley, 1950.

Kaplan, H. Prediction of Success in Army Aviation Training. APRO Technical Research Report 1142, June 1965.

Kaplan, H. Psychological Testing Programs in the US Army. ARI Research Report, September 1973.

Maier, M. H. Procedures for Assigning Criterion Grades to Failures and Turnbacks in Army School Courses. USABESRL Technical Research Nore 197, April 1968.

Rosenberg, N., Martinek, H. and Anderson, A. A. Development of a Provisional Battery for Selecting Army Helicopter Pilots. PRB Technical Research Note 104, June 1959.

MINUTES

1975 MTA Steering Committee Meeting

The 1975 Military Testing Association (MTA) Steering Committee Meeting opened at 4:00 P.M. at the Marriott Inn, Indianapolis, Indiana on 15 September 1975.

The following conferees were present:

U. S. Army Enlisted Evaluation Center (Coordinating Agency)

COL George W. Orton, President
Dr. Clay V. Brittain, Conference Chairman
Mr. William Wance

U. S. Armed Forces Vocational Testing Group

COL R. S. Hoggatt, USAF
Dr. Harry D. Wilfong

U. S. Naval Education & Training Program Development Center

Captain Richard T. Thomas
Mr. Walter W. Birdsall

U. S. Naval Personnel Research & Development Center

Mr. Martin Wiskoff

U. S. Coast Guard Institute

LCDR Raymond D. Bland
Mr. John Burt

U. S. Marine Corps Headquarters

LTC H. L. Angle
LTC David Vowell

U. S. Air Force Human Resources Laboratory

Dr. Raymond E. Christal

Royal Military College of Canada

L. Col Gerald J. Carpenter

Royal Australian Air Force

Squadron Leader Kenneth Goody

U. S. Civil Service Commission, Chicago

Mr. Cas Winiewicz

Training Management Institute, TRADOC

CPT Harry W. Porthouse

COL Orton, President of the MTA, opened the meeting and Dr. Brittain presented the agenda items previously submitted by members of the steering committee.

1. The first item on the agenda was consideration of a name change of the Association. It was noted that this had been a recurring issue over the past several years. Discussion of merits of a name change surfaced the fact that some agencies desiring to participate in the MTA have difficulty justifying their participation because of the limited focus implied in the name of the Association. It was suggested that the charter of the Association was broad enough in content to cover those agencies. A motion was made by Dr. Christal that the name be changed to Military Evaluation Association. The motion was seconded but failed to carry by majority vote.

2. A discussion followed on whether a change in the statement of the purpose of MTA as found in the By-Laws would be beneficial in emphasizing the scope of the Association. Captain Thomas made a motion to rearrange the sections under Article II, with section C becoming section A. Motion was lost without a second.

3. MTA Symbol -- In the past the symbol has varied from year to year. It was suggested that the symbol be standardized. Motion was made by Colonel Carpenter that the standard logo be adopted and used in the center without the outside circle. Motion was seconded and carried unanimously.

4. Service plans for validating new Armed Services Vocational Aptitude Battery (ASVAB). Dr. Harry Wilfong suggested that in view of the service-wide implications associated with January 1976 implementation of the revised ASVAB, that service plans for validating the new battery be addressed. It was generally agreed that the MTA steering committee was not a suitable forum to discuss this topic.

5. Rotating Sponsorship of Annual Conferences -- The Marine Corps has withdrawn as a sponsor of the MTA annual conference. As a result of the withdrawl, changes to "B" under Article VII were discussed, It was suggested that sponsorship be by Agency rather than services, however, in consideration of other impending changes within several commands of the Armed Services, it was agreed to table until next year the question of changes in the By-Laws on the sponsorship of annual MTA conferences. The Navy will host the MTA in 1976. The Air Force was tentatively identified as the host in 1977.

6. Nominations for the Harry H. Greer Award -- Two issues were raised in regard to the Harry H. Greer Award. (a) The handling of nominations for the Award this year, and (b) the establishment of selection criteria and procedures for future years. Following discussion by the steering committee, COL Orton appointed a committee to develop recommendations in regard to both issues. The committee is to be made up of one person from each of the following agencies: US Army Enlisted Evaluation Center, US Coast Guard Institute, US Air Force Human Resources Laboratory, US Armed Forces Vocational Testing Group, US Naval Education & Training Program Development Center, and the US Naval Personnel Research & Development Center. The member from USAEEC will serve as committee chairman.

7. It was suggested that the Army Research Institute for the Social and Behavioral Sciences be invited to membership in the MTA Steering Committee. COL Orton spoke in favor of ARI becoming a member of the steering committee since it has a large representation at the MTA. Mr. Burt made a motion that the MTA extend an invitation to the Army Research Institute to primary membership in MTA. The motion was seconded and was carried unanimously.

8. Mr. Cas Winiewicz requested that the U. S. Civil Service Commission be invited to membership in the MTA. Mr. John Burt made a motion that the U. S. Civil Service Commission be invited to associate membership in MTA. The motion was seconded and carried unanimously. The letter of invitation will be sent to the Chairman, U. S. Civil Service Commission.

9. The membership status of the Marine Corps was discussed. Since the Marine Corps can no longer sponsor annual MTA Conferences, should the Marine Corps membership be converted from primary to associate membership? John Burt made a motion that the Marine Corps membership in MTA be associate membership. The motion was seconded and carried unanimously. The Marine Corps representatives concurred in the appropriateness of this action. A motion was made by Squadron Leader Goody that the types of membership on the steering committee be resolved at next years MTA conference. The motion was seconded and passed unanimously.

10. The membership status of other agencies was discussed. It was generally agreed that primary membership should be reserved for those agencies which sponsor annual MTA conferences, with other agencies having associate membership.

The Steering Committee reconvened at 1450, 18 September.

11. The Harry H. Greer Award

a. The committee appointed by Colonel Orton reported that there were no nominations for the award this year.

b. The committee proposed that the MTA By-Laws be amended to incorporate the Article at Inclosure 1. John Burt made a motion that this proposal be adopted. Colonel Hoggatt seconded the motion and it passed unanimously.

12. Dr. Christal made a motion to extend to the Air Force Occupational Measurement Center an invitation to MTA membership. Mr. Birdsall seconded the motion and it was passed unanimously.

Proposed Amendment to MTA By-Laws

Article XII

Harry H. Greer Award

A. Selection Procedures:

1. Recipients of the Harry H. Greer award will be selected by a committee drawn from the agencies represented on the MTA Steering Committee. The CO of each agency will designate one person from that agency to serve on the awards committee. Each committee member will have attended at least three previous MTA meetings. The member from the coordinating agency will serve as chairman of the committee.

2. Nominations for the award in a given year will be submitted in writing to the awards committee chairman by 1 January of that year.

3. The chairman of the committee is responsible for canvassing the other committee members to arrive at concensus on the selection of a recipient of the award.

4. Not more than one person is to receive the award each year, but the award need not be made each year. The Awards Committee may decide not to select a recipient in any given year.

5. The annual selection of the person to receive the award or the decision not to make an award that year is to be made at least six weeks prior to the date of the annual MTA Conference.

B. Selection Criteria:

The recipients of the Harry H. Greer award are to be selected on the basis of outstanding work contributing significantly to the MTA.

C. The Award:

The Harry H. Greer Award is to be a certificate normally presented to the recipient during the Annual MTA Conference. The awards committee is responsible for preparing the text of the certificate. The coordinating agency is responsible for printing and awarding the certificate.

Inclosure 1

Minutes of Concluding General Session of the MTA Conference

The meeting was called to order at 1100, 19 September 1975. COL Orton, MTA President presided.

1. Copies of the minutes of the Steering Committee were distributed to the assembled members. Colonel Orton asked for discussion of the minutes, especially of the proposal that the MTA By-Laws be amended by the adding of Article XII pertaining to the Harry H. Greer Award.

Major Warren B. Shull made a motion that the proposed amendment be adopted. John Finucane seconded the motion and it was passed unanimously.

2. Colonel Orton announced that Dr. William Moonan of the Navy Personnel Research and Development Center, San Diego, is to develop a form to be used in rating nominees for the Greer Award. Dr. Raymond Christal noted that in developing factors for this form, Dr. Moonan wishes to get suggestions from MTA members. Suggestions should be sent directly to Dr. Moonan.

3. Colonel Orton mentioned that the matter of changing the name of MTA was discussed by the Steering Committee, but the motion on a specific name change was defeated by a vote of the Steering Committee. The Navy Education and Training Program Development Center, as coordinator of the 1976 MTA Conference, has indicated willingness to canvass MTA members to get reactions to a number of possible names. The three names most frequently favored could be put to a floor vote at the 1976 conference.

4. LTC Peter E. LaSota of USAF Occupational Measurement Center expressed appreciation that the AF Occupational Measurement Center was to be invited to membership in MTA.

5. Dr. Joseph Boyd of the Education Testing Service, Princeton, NJ, reported on his discussions with the ERIC Clearinghouse on Educational Tests and Measurements concerning the inclusion of the MTA Proceedings in the ERIC System. The Clearinghouse is to abstract the proceedings as a whole and enter the abstract into the ERIC System. Additionally, selected papers will be entered in unabridged form. Copies of the abstract of the proceedings as a whole and of the selected papers will be available through ERIC in either microfische or hard copy. In order that proceedings of conferences prior to 1974 be entered into ERIC, Dr. Boyd requires two copies of the proceedings. The coordinating agencies for these conferences should send copies of the proceedings to Dr. Joseph Boyd, Education Testing Service, Princeton, New Jersey 08540.

6. The 1976 MTA conference will be hosted by the Navy Education and Training Program Development Center, Pensacola, Florida. The conference is scheduled for the first week of November, 1976.

7. The final session of the 1975 conference adjourned at 1120.

BY-LAWS OF THE MILITARY TESTING ASSOCIATION

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of various governmental, educational, business, industrial and private organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems. Each agency shall have no more than two (2) professional civilian representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental, educational, business, industrial and private organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of the Association or the Chairman of the Steering Committee. A majority of the members of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment in order of the following rotating schedule:

United States Army
United States Marine Corps
United States Navy
United States Air Force
United States Coast Guard

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be an ex officio member of all standing committees.

B. The President with the counsel and approval of the Steering Committee may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Enactment

These By-Laws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 2 November 1973).

Article XII - Harry H. Greer Award

A. Selection Procedures:

1. Recipients of the Harry H. Greer award will be selected by a committee drawn from the agencies represented on the MTA Steering Committee. The CO of each agency will designate one person from that agency to serve on the awards committee. Each committee member will have attended at least three previous MTA meetings. The member from the coordinating agency will serve as chairman of the committee.

2. Nominations for the award in a given year will be submitted in writing to the awards committee chairman by 1 January of that year.

3. The chairman of the committee is responsible for canvassing the other committee members to arrive at concensus on the selection of a recipient of the award.

4. Not more than one person is to receive the award each year, but the award need not be made each year. The Awards Committee may decide not to seleet a recipient in any given year.

5. The annual selection of the person to receive the award or the decision not to make an award that year is to be made at least six weeks prior to the date of the annual MTA Conference.

B. Selection Criteria:

The recipients of the Harry H. Greer Award are to be selected on the basis of outstanding work contributing significantly to the MTA.

C. The Award:

The Harry H. Greer Award is to be a certificate normally presented to the recipient during the Annual MTA Conference. The awards committee is responsible for preparing the text of the certificate. The coordinating agency is responsible for printing and awarding the certificate.

LIST OF CONFEREES

ANGLE, LTC H. L.
HQ, US Marine Corps, Office of Manpower Utilization, Quantico, VA 22134
ANSBRO, Mr. Thomas M.
OPO OFIT, Ellyson Field, Pensacola, FL 32509
APPLEGATE, Mr. Neil J.
Naval Education & Training Support, Ellyson Field, Pensacola, FL 32509
ARNOLD, Mr. George E.
Test Project Director, USA Armor School, Fort Knox, KY
BACHTEL, Ms Mary A.
Naval Education & Training Program Development Center (PDD5),
Pensacola, FL 32509
BAREFIELD, Ms Ruth B.
Department Army-Wide Training Support, US Army Aviation Center,
Ft. Rucker, AL 36362
BARNEY, Mr. Douglas C.
Design Division, US Army Institute of Administration, Ft. Harrison, IN 46216
BASKIN, Mr. Wallace N.
US Army Missile & Munitions Center and School (ATSK-CTD-DP),
Redstone Arsenal, AL 35809
BAYMOR, Mr. Michael
Evaluation Branch, US Army Armor School, Ft. Knox, KY 40121
BEGLAND, CPT Robert
Faculty Development Division, US Army Infantry School, Ft. Benning, GA 31905
BERGER, Mr. B. Michael
US Army Military Occupational Research Division, Alexandria, VA 22331
BIRDSALL, Mr. Walter W.
Naval Education & Training Program Development Center (PDE-E),
Ellyson Field, Pensacola, FL 32509
BLAND, LCDR Raymond D.
U.S. Coast Guard Institute, P. O. Substation 18, Oklahoma City, OK 73169
BOWSER, Mr. Samuel E.
Naval Personnel Research and Development Center (Code 301),
San Diego, CA 92152
BOUZA, Mr. Duane
Guidance Counselor, Base Education Center (DPT), Offutt AFB, NE 68113
BOYD, Dr. J. L.
Educational Testing Service, Princeton, NJ 08540
BOYD, Mr. Richard D.
US Army Military Occupational Research Division, Alexandria, VA 22331
BRADDOCK, Mr. Edward
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
BRAND, Mr. John S.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
BRANSON, Dr. Robert R.
Florida State University, 1A Tully Gym, Tallahassee, FL 32306

BRASHEAR, Mr. Fenton W.
US Army Ordnance Center & School, Aberdeen Proving Ground, MD 21005
BRAUN, Mr. Frank B.
Data-Design Laboratories, 15 Koger Executive Center, Suite 140,
Norfolk, VA 23502
BRITTAIN, Dr. Clay V.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
BROGA, LT Dwight C., III
U. S. Coast Guard (G-PE-1/72), 400 Seventh Street, SW, Washington, DC 20590
BROMAN, Mr. James D.
Assistant Director Examination Programs, Defense Activity for Non-Traditional Education Support (DANTES), Ellyson Field, Pensacola, FL 32509
BRONSTEIN, Mr. Russell H.
US Army Institute of Administration, Ft. Harrison, IN 46216
BROWN, CDR Bruce W.
CNET Staff, Naval Air Station, Pensacola, FL 32508
BUCK, Dr. C. W.
U. S. Coast Guard Institute, P. O. Substation 18, Oklahoma City, OK 73169
BUCK, Mr. Lawrence S.
US Civil Service Commission, 19th & E Streets, NW, Washington, DC 20415
BURNSIDE, CPT Billy L.
TAGCEN HQDA, (DAAG-SD), Forrestal Bldg. Rm 5B138, Washington, DC 20314
BURT, Mr. John A.
U. S. Coast Guard Institute, P. O. Substation 18, Oklahoma City, OK 73169
BUSH, CWO3 B. J.
U. S. Coast Guard Training Center, Petaluma, CA 94952
CALLAHAN, Mr. M. D.
Technical Director, Navy Occupational Task Analysis Program (NOTAP),
Bldg 1304, Bolling AFB, Washington, DC 20332
CANTER, Dr. Ralph R.
US Army Research Institute, 1300 Wilson Boulevard, Arlington, VA 22209
CARDACI, LCDR Philip J.
U. S. Coast Guard (G-PTE-3/72), 400 Seventh Street, SW, Washington, DC 20590
CARPENTER, LTC Gerald J.
Royal Military College of Canada, Kingston, Ontario, Canada K7L2W3
CAVINESS, Dr. J. A.
US Army Research Institute, P. O. Box 2086, Ft. Benning, GA 31905
CHIPPENDALE, Ms Joan T.
U. S. Coast Guard Training Center, Governors Island, New York, NY
CHRISTAL, Dr. Raymond E.
AFHRL, Lackland AFB, TX 78148
COCKRELL, Dr. John T.
US Army Research Institute, Ft. Knox, KY 40121
CURL, MAJ William W., III
US Army Combat Arms Training Board, Ft. Benning, GA 31905
DAMON, LTC James A.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249

DAVIS, Mr. D. D.
Naval Education & Training Program Development Center, Ellyson Field,
Pensacola, FL 32509
DeVEAU, Mr. Larry E.
Navy Technical Training Command (Code 0163), NAS Memphis, Millington, TN 58034
DRISKILL, Dr. Walter E.
USAF Occupational Measurement Center, San Antonio, TX
DYGERT, LTC George H.
The Judge Advocate Generals' School, Charlottesville, VA 22901
EDWARDS, Mr. Richard K.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
EVANS, Mr. E. M.
CNTECHTRIT (Code 01622), Millington, TN
FELKINS, Ms Deborah A.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
FINUCANE, Mr. John L.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
FISCHL, Dr. M. A.
US Army Research Institute, Washington, DC
FITES, Ms Jeanne B.
Manpower Research & Data Analysis Center, 300 N. Washington Street,
Alexandria, VA 22314
FOLEY, Mr. Bernard J.
US Army Security Agency Training Center & School, Ft. Devens, MA 01433
FULLER, CAPT John H.
1051 S. Dobson, #89, Mesa, Arizona 85202
GARY, Mr. Perry L.
US Army Aviation School, Ft. Rucker, AL 36362
GEDDIE, Dr. James C.
US Army Research Institute for the Behavioral and Social Sciences
ARI Field Unit, Ft. Hood, TX 76544
GEESEY, Mr. Scott S.
US Army Quartermaster School, (ATSM-CTD-TA), Ft. Lee, VA 23801
GERBER, Mr. J. E. Jr.
Education Services Officer, US Army Forces Command, Ft. McPherson, GA 30330
GIERHART, Mr. Harry S.
U. S. Coast Guard, (G-P01/62), Washington, DC 20590
GILBERT, Dr. Arthur C. F.
US Army Research Institute, Arlington, VA 22209
GLENN, MAJ John T.
Office of the Deputy Chief of Staff for Personnel, DAPE-MPE-CS,
Washington, DC 20310
GLENN, Mr. Travis C.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
GOODEN, MAJ James R.
US Army Ordnance Center & School, Aberdeen Proving Ground, MD 21001
GOODY, Squadron Leader Kenneth
Royal Australian Air Force, AFHRL/OR, Lackland AFB, TX 78236
GOULD, Mr. R. Bruce
AFHRL/ORE, Lackland AFB, TX 78236

GMEINDER, Dr. Susan T.
135 Sea Foam Avenue, Monterey, CA 93940
GREGORY, CAPT Robert A.
6613 Potomac Avenue, Apt. A-1, Alexandria, VA 22307
GROSCOST, Mr. John
National Security Agency, Ft. Meade, MD 20755
GUIBERSON, MAJ Ronald M.
US Army Institute of Administration, Ft. Harrison, IN 46216
HALTRECHT, Dr. Ed
Personnel Research, Ontario Hydro, 620 University Avenue, Toronto, Canada
HAMMER, Ms Kathleen A.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
HANNIGAN, Mr. Matthew
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
HARMAN, Mr. Samuel A.
US Army Facilities Engineering Support Agency, Ft. Belvoir, VA 22060
HAWKINS, Mr. Helmut H.
US Army Military Occupational Research Division, Alexandria, VA 22331
HAZEL, Dr. Joe T.
Occupational Research Division, AFHRL, Lackland AFB, TX 78236
HECKINGER, Ms Amy R.
US Army Intelligence Center & School, Ft. Huachuca, AZ 85613
HERMANSEN, Mr. Arthur G.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
HILL, Mr. Norman E.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
HISCOX, Mr. Michael D.
Northwest Regional Education Laboratory, 710 SW 2nd Avenue, Portland, OR 97204
HOGGATT, COL Ralph S.
Armed Forces Vocational Testing Group, AFVTG/CC, Randolph AFB, TX 78148
HOGGE, CAPT Robert M.
Extension Course Institute, Gunter AFS, AL 36118
HOHREITER, Mr. Jack E.
6516 N. Ferguson Avenue, Indianapolis, IN 46220
HONN, Mr. Fred B.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
HORAN, LT(JG) W. T.
U. S. Coast Guard Headquarters, Washington, DC 20590
HOYT, Dr. William G.
System Development Corporation, 2500 Colorado Avenue, Santa Monica, CA 90406
HUDDLESTON, Mr. James E., Jr.
US Army Training and Doctrine Command, ATTNG-AS-IT, Ft. Monroe, VA 23651
JACKSON, Mr. William L.
US Army Aviation Center, Ft. Rucker, AL 36362
JONES, Ms Anita
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
JOHNSON, Mr. Otis J.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
JOHNSON, Mr. Robert N.
US Army Institute of Administration, Ft. Harrison, IN 46220

JOHNSON, LTC Russell W.
Occupational Measurement Center, Lackland AFB, TX 78236
KACZMARSKI, Mr. John F.
US Army Infantry School, Ft. Benning, GA 31905
KANTOR, Dr. Jeffrey E.
AFHRL, Lackland AFB, San Antonio, TX 78236
KELLEHER, Ms Anna May
Con Edison New York, 4 Irving Place, New York, NY 10003
KESSLER, Dr. John J.
US Army Research Institute, Arlington, VA 22209
KING, Professor F. J.
Combat Arms Training Board, 405 Education, Florida State University,
Tallahassee, FL 32306
KNIGHT, MAJ Daniel J.
US Army Security Agency, Arlington Hall, Arlington, VA 22212
KOHL, Mr. Delbert E.
Marine Corps Institute, Box 1775, Washington, DC 20013
KOYM, Mr. Kenneth
AFHRL, Lackland AFB, TX 78236
KRAIN, Dr. Burton
US Civil Service Commission, 230 S. Dearborn, Chicago, IL 60604
KREIGER, Mr. Gregory M.
33 Highland Avenue, Leominster, MA 01453
LaSOTA, LTC Peter E.
USAF Occupational Measurement Center, Lackland AFB, TX 78236
LANG, Ms Darryl A.
U. S. Coast Guard Headquarters (G-P-1/62), Washington, DC 20590
LANTERMAN, Mr. Richard S.
U. S. Coast Guard Headquarters (G-P-1/62), Washington, DC 20590
LEESON, LTC Peter
Embassy of Australia, Army Staff, 1601 Massachusetts Avenue,
Washington, DC 20036
LONG, Dr. George E.
McDonnell Douglas Corporation, St. Louis, MO
McBRIDE, Mr. Curtis D.
US Army Field Artillery School, St. Sill, OK
McDONALD, Mr. John D.
US Army Institute of Administration, Ft. Harrison, IN 46216
McGUINNESS, Dr. James
ESSEX Corporation, 201 North Fairfax Street, Alexandria, VA 22314
McMULLEN, Mr. Robert L.
US Army Research Institute Field Unit, Ft. Rucker, AL 36362
McNICHOLS, MAJ Chuck
Air Force Institute of Technology, Wright-Patterson AFB, OH
MAHNEN, Mr. Herman A.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
MANLEY, LTC T. R.
Air Force Institute of Technology, Wright-Patterson AFB, OH

MATTHEWS, Dr. Gary N.
Air Force Human Resources Laboratory, Lackland AFB, TX
MEAD, LTC Donald F.
Air Force Human Resources Laboratory, Lackland AFB, TX
MEEHAN, Dr. John T.
Headquarters Air University, AU/EDV, Maxwell AFB, AL 36112
MILLER, CDR James P.
Executive Agent's Rep (for DOD) for Joint Task Analysis Support,
NOTAP (NPPSA), Building 1304, Bolling AFB, Washington, DC 20332
MITCHELL, MAJ Jimmy L.
2719 Henderson, West Lafayette, IN 47906
MOONAN, Dr. William J.
US Navy Personnel Research and Development Center, San Diego, CA 92152
MORHARDT, Ms Francia G.
U. S. Postal Service, 475 L'Enfant Plaza West, SW, Washington, DC 20260
MORTON, LTC Bryan D.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
NAFZIGER, Mr. Dean H.
Northwest Regional Educational Laboratory, 710 SW Second Avenue,
Portland, OR 97204
NEAL, CPT(P) William D.
Combat Arms Training Board, TEC Division, Ft. Benning, GA 31905
NEFF, Mr. Edwin F.,
U. S. Coast Guard (G-PTE/72), 400 7th Street, SW, Washington DC 20590
NICHOLL, CDR C. J.
Naval Manpower Utilisation Unit, HMS Vernon, Portsmouth, England
NOWE, Mr. C. Edward
Federal Aviation Administration, FAA Aeronautical Center, P. O. Box 25082,
AAC 234, Oklahoma City, OK 73125
O'CONNOR, MAJ Thomas J.
USAF Occupational Measurement Center, Lackland AFB, TX
OCCHIALINI, Mr. Chris
PACDA-HRD, Ft. Harrison, IN
OHLE, CPT David H.
U. S. Military Academy, Office of Military Leadership, West Point, NY 1099
ORTON, COL George W.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
PARKS, Mr. Cortez
Defense Language Institute, DLIELC (English Language Center),
Lackland AFB, TX 78236
PLANERT, Ms. Diane E.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
PORTER, Mr. George V., Jr.
US Air Force Academy, USAFA/RRR, USAF Academy, CO 80840
PORTHOUSE, CPT Harry W.
Training Management Institute, US Army TRADOC, Attn: ATTNG-TMI,
Ft. Eustis, VA 23604
PRESTRIDGE, MSG William C.
US Army Facilities Engineering Support Agency, Ft. Belvoir, VA 22060
PRIM, MAJ David F.
US Army Military Police School, DCET, Ft. McClellan, AL 36201

RAFACZ, Mr. Bernard A.
Navy Personnel Research and Development Center, San Diego, CA 92152
RAINES, Mr. John T.
6001 Frederick Street, Springfield, VA 22150
RAMSEY, Mr. Wendell N.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
RAYNER, Dr. Gail T.
Florida State University, 1A Tully Gym, Tallahassee, FL 32306
REED, Ms Janice A.
US Army Institute of Administration, Ft. Harrison, IN 46216
RIGBY, Ms Martha J.
US Army Infantry School, Ft. Benning, GA 31905
ROBERTS, Mr. Charles D.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
ROBERTS, **CPT Wesley K.**
US Army Combat Arms Training Board, Ft. Benning, GA
ROBINSON, Dr. M. A.
Data-Design Laboratories, 15 Koger Executive Center, Suite 140,
Norfolk, VA 23502
ROSE, Mr. LeRoy
US Army Transportation School, Ft. Eustis, VA 23604
RUBAK, Dr. Jack
Defense Information School, Ft. Harrison, IN
RUCK, Mr. Hendrick W.
USAF Occupational Measurement Center, Lackland AFB, TX 78236
SACHSE, Mr. Thomas P.
Clearinghouse for Applied Performance Testing, Northwest Regional
Educational Laboratory, 710 SW Second Avenue, Portland, OR 97204
SANDS, Mr. William A.
Navy Personnel Research and Development Center, San Diego, CA 92152
SCANLAND, Dr. Francis W.
Chief of Naval Education and Training (Code N-5A), Naval Air Station,
Pensacola, FL 32508
SCHROEDER, Mr. Paul E.
Ohio State University, Center for Vocational Education, 1960 Kenny Road
Columbus, OH 43210
SCHULTZ, **SP7 William E.**
USAE, SOM, USN Anphibase, Norfolk, VA 23521
SCHUMACHER, CPT Daniel L.
US Army Institute of Administration, Ft. Harrison, IN 46216
SCHUSTER, Mr. Bernie
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
SHULL, MAJ Warren B.
Air Force Reserve Officer Training Corps, Maxwell AFB, AL 36112
SMITH, Mr. Carl R.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
SMITH, Dr. Robert G., Jr.
7606 Chancellor Way, Springfield, VA 22153
SMITH, Mr. Ronald H.
Naval Education and Training Program Development Center, Ellyson Field,
Pensacola, FL

SNELLING, MSG Lonie E.
US Army Chaplain Center and School, Ft Wadsworth, Staten Island, NY 10305
SNYDER, LTC John S.
US Army Chaplain Center and School, Ft. Wadsworth, Staten Island, NY 10305
SPECTOR, Mr. Daniel E.
US Army Military Police School, Ft. McClellan, AL 36201
STACY, Mr. William J.
Air Force Human Research Laboratory, Lackland AFB, TX
STANDLEE, Dr. Lloyd S.
Navy Personnel Research and Development Center, San Diego, CA 92152
STEARNS, SSG Ralph I.
3902nd ABWg/DPT, Offutt AFB, NE 68113
STEINHEISER, Dr. Frederick
US Army Research Institute, Arlington, VA 22209
STRIGARI, Mr. Anthony J.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
STRUB, Dr. Michael H.
US Army Research Institute Field Unit, Building 118, Ft. Bliss, TX 79916
SWANSON, MAJ Ronald G.
US Air Force Academic Instructor School, AIAOS/EDB, Maxwell AFB, AL 36112
SZASZY, Mr. Alex L.
Defense Language Institute, P.O. Box 5794, Monterey, CA 93940
TAKAHASHI, Mr. Terry Y.
Defense Intelligence School, Washington, DC 20374
TINDALL, Mr. J. E., Jr.
Data-Design Laboratories, 15 Koger Executive Center, Suite 140,
Norfolk, VA 23502
THOMAS, Dr. David B.
University of Iowa: CAI Laboratory, 229 Lindquist Center for Measurement,
Iowa City, IA 52242
THOMAS, CPT R. T.
Naval Education and Training Program Development Center, Ellyson Field,
Pensacola, FL 32509
THURING, LT(JG) Allen R.
US Coast Guard, (G-RT-2/81), Washington, DC 20590
TRATHNER, Dr. Marvin H.
U. S. Civil Service Commission, 1900 E. Street, NW, Washington, DC 20415
TURLEY, LCDR L. A.
Royal Australian Navy, CNET, Code N-534, Naval Air Station, Pensacola FL 32508
TUTTLE, Dr. Thomas C.
Westinghouse Behavioral Services, Box 948, American City Building,
Columbia, MD 210441
UBERTA, Ms Linda
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
UNDERWOOD, MAJ Edward R.
Department of Army Wide Training Support, Ft. Rucker, AL 36362
VAN CLEVE, COL Roy R. (Ret)
Director of Personnel, Comptroller of Public Accounts, State of Texas,
Room 119, LBJ Building, Austin, TX

VALENTINE, Dr. Lonnie D.
Air Force Human Resources Laboratory, Lackland AFB, TX 78236
VAUGHAN, 1LT David S.
USAF Occupational Measurement Center, Lackland AFB, TX
VOORHEES, Ms Phyllis
U. S. Coast Guard Institute, Substation 18, Oklahoma City, OK
VOWELL, LTC David E.
Office of Manpower Utilization, U. S. Marine Corps Headquarters,
(Code MPU) Marine Corps Base, Quantico, BA 22134
WALDKOETTER, Dr. Ray
U. S. Army Research Institute, Arlington, VA 22209
WALKER, LCDR Lee
Central Test Site for Personnel and Training Evaluation Program (PTEP)
NAUGMS, Dam Neck, VA 23641
WALSH, Mr. Lawrence A.
Publisher, Career Education, McGraw-Hill Book Company, 1221 Avenue
of the Americas, New York, NY 10020
WANCE, Mr. W. W.
US Army Enlisted Evaluation Center, Ft. Harrison, IN 46249
WASER, Mr. Charles D.
U. S. Coast Guard Headquarters (G-P-1/62), Washington, DC 20590
WEISS, Professor David J.
University of Minnesota, N660 Elliott Hall, Minneapolis, MN 55455
WEISSMULLER, Mr. Johnny J.
Air Force Human Resources Laboratory, Lackland AFB, TX
WETHY, Mr. Roy B.
US Army Military Occupational Research Division, Alexandria, VA 22331
WILFONG, Dr. Harry D.
Armed Forces Vocational Testing Group, AFVTG/RD, Randolph AFB, TX 78148
WILLIAMS, Mr. Rayburn A.
Chief of Naval Education and Training, Pensacola, FL 32506
WILLING, Mr. Richard C.
U. S. Coast Guard Institute, P. O. Substation 18, Oklahoma City, OK 73132
WINIEWICZ, Mr. Cas
U. S. Civil Service Commission, 230 S. Dearborn Street, Chicago, IL 60604
WISKOFF, Dr. Martin F.
Navy Personnel Research and Development Center, San Diego, CA 92152
WOOD, CDR B. L.
U. S. Coast Guard Training Center, Petaluma, CA 94952
WRIGHT, MSG Lyle A.
US Army Institute of Administration, Ft. Harrison, IN 46216
YELLEN, Mr. Ted
Navy Personnel Research and Development Center, San Diego, CA 92152
YORK, Mr. William J.
US Army Signal School, Ft. Gordon, GA

☆U.S.GOVERNMENT PRINTING OFFICE:1976 -653-757